Investigating the role of aberrant gene regulation in inflammatory bowel disease to understand pathogenesis and help predict relapse

Demandt, Sanne Laura Jacqueline

Awarding institution:
King's College London

Investigating the role of aberrant gene regulation in inflammatory bowel disease to understand pathogenesis and help predict relapse

by

Sanne Laura Jacqueline Demandt

A thesis submitted to the University of London in fulfilment of the degree of Doctor of Philosophy

King's College London

Department of Medical and Molecular Genetics

Division of Genetics and Molecular Medicine

September 2017

Abstract

IBD is a complex genetic disease characterised by chronic inflammation of the gastro-intestinal tract, with Crohn's disease (CD) and ulcerative colitis (UC) being the two most common forms. Genome wide association studies (GWAS) and meta-analysis have identified >200 genomic IBD susceptibility regions, the most of any complex disease. Here, whole transcriptome sequencing was employed to investigate the role of altered gene expression in intestinal tissue.

Differential expression and potential underlying biological pathways were assessed between UC, CD and controls. Furthermore, the effect of indexed IBD risk SNPs on changes in gene expression was investigated. Heterogeneity within the RNA sequenced intestinal biopsy samples was addressed through cellular phenotyping and computational sample deconvolution. Additionally, the presence of a transcriptional signature to predict relapse was investigated.

1,637 transcripts exhibited differential expression at $q \leq 0.05$ between the IBD sub phenotypes and controls. Most notably, *GLS* (Glutaminase), an enzyme which hydrolysis glutamine into glutamate and ammonia. Glutamine is known to be an important energy source for immune and gut mucosal cells. Furthermore, it was observed that these differentially expressed genes significantly perturbed 50 biological pathways. The majority of the identified pathways were involved in processes known to play an important role in IBD: immune regulatory, autophagy and transmembrane signalling. One novel finding was the perturbation of Nicotine degradation pathway II and III within CD patients *versus* controls and UC patients. Potentially providing insight into the mechanism behind the known opposing effects of smoking on the clinical course of UC and CD patients. Expression of 9 genes located within an IBD loci showed association with an IBD risk SNPs, making them strong candidate genes in IBD pathogenesis and further investigation should be performed.

Acknowledgements

Dr Natalie Prescott, my primary supervisor, I want to thank for giving me the opportunity to undertake this project and for her supervision throughout the 3 years of this PhD. Furthermore, Dr Peter Irving, my second and clinical supervisor, my gratitude for introducing me to the clinical aspects of the project. Without his guidance and assistance, the samples collection and thus the project would not have been possible. Furthermore, my appreciation to Professor Christopher Mathew, for his contribution to the overall process.

Many thanks to the rest of the Complex Disease Group for making my time enjoyable: Phillip Tombleson, for his never failing willingness to help, Yasmin Omar, for never letting us have a dull moment, and foremost Ariella Amar, for all her hard work and assistance with experiments but even more so for being my friend and supporting me through the harder times of the PhD.

I also thank and acknowledge the members of the Gastroenterology team, both at Guy's hospital and St Thomas' hospital, without whom I would have never hit my sample collection goals. All the registrars for their help in collecting biopsies, the nursing staff for aiding me where needed and particularly Dr Kamal Patel, for collecting all the post-surgery patient samples.

Many thanks to the BRC Bioinformatics unit for their collaboration on the RNA sequencing data analysis. Especially Dr Philipe Gracia, for not only performing the analysis but taking the time to teach me and provide me with the tools to perform the analysis myself.

My gratitude to the staff at the BRC Genomics Core facility for their help with the RNA sequencing, genotyping and processing of the microarrays. Furthermore, I want to express my appreciation for their willingness to give advice and a helping hand whenever needed.

I want to say thank you to the BRC Flow Cytometry Core for cell sorting the relapse samples and their input in optimising the biopsy phenotype antibody staining.

Ken Hanscombe and Seth Seegobin, I want to thank for their assistance with my statistical analysis. With a special mention for Seth, for not only helping on a scientific level but for being my friend.

I am grateful to the Biomedical Research Centre (BRC) for funding my PhD. The King's Bioscience Institute (KBI) for co-organising the PhD programme. In addition, a big thank you to all my fellow BRC/KBI PhD student for making this journey sociable and fun.

Finally, I want to express my gratitude and appreciation to my family. My mom and dad, Liny and Peter, for always being supportive, giving advice and believing in me. My sister, Lune, for always offering a listing ear and being able to cheer me up. Most of all I am thankful to Sam, my boyfriend, for his contributions to my thesis and more so for putting up with the long hours, the stress levels and my bad moods. You always knew what to say to make it better.

Table of contents

List of Figures

List of Tables

List of Appendices

List of Abbreviations

Abbreviation	Meaning
ΔCt	delta-Ct
anti-TNF	anti-Tumour Necrosis Factor
ASA	Aminosalicylates
ASCA	anti-*Saccharomyces cerevisiae* antibody
CARD	Centre for Age Related Disease
CD	Crohn's disease
CMP	counts per million
CRP	C-reactive protein
Ct	Cycle threshold
DAPPLE	Disease Associated Protein-Protein Link Evaluator
DCs	Dendritic cells
DE	Differential expression
DNA	Deoxyribonucleic acid
DTPA	Diethylenetriamine Pentaacetate
E.coli	*Escherichia coli*
eQTL	expression Quantitative Trait Loci
ERCC	External RNA Control Consortium
ES	Enrichment score
ESR	Erythrocyte sedimentation rate
FDR	False discovery rate
FPKM	Fragment per kilobase of exon per million fragments mapped
FT	Transcription factor
GI	Gastrointestinal tract
GRAIL	Gene Relationship Across Implicated Loci
GSEA	Gene Set Enrichment Analysis
GTEx	Genotype-Tissue Expression project
GWAS	Genome Wide Association Study
IBD	Inflammatory bowel disease
IFNa	Interferon-alpha
IFN-y	Interferon-gamma
IPA	Ingenuity Pathway analysis
LD	Linkage disequilibrium
lncRNA	Long non-coding RNA
LPS	Lipopolysaccharide
MSigDB	Molecular Signature Database
MZ	Monozygotic
NES	Normalised enrichment score
NF-$_k$B	Nuclear factor-$_k$B
NGS	Next generation sequencing

NK	Natural killer
nt	Nucleotides
pANCA	perinuclear Anti-Neutrophil Cytoplasmic Antibody
PBMCs	Peripheral blood mononuclear cells
PC	Principle component
PCA	Principle components analysis
PRR	Pattern recognition receptor
QQ	Quantile-Quantile
RA	Rheumatoid arthritis
RIN	RNA integrity score
RNA	Ribonucleic acid
RQ	Relative quantification
rRNA	ribosomal RNA
RT	Reverse transcriptase
RT-PCR	Reverse transcription polymerase chain reaction
RUV	Remove unwanted variables
SNPs	Single nucleotide polymorphisms
sscDNA	single stranded cDNA
TLR	Toll-like receptor
TMM	Trimmed mean of M
Tregs	T regulatory cells
UC	Ulcerative colitis
WTCCC	Wellcome Trust Case Control Consortium

1. Introduction

1.1 Inflammatory bowel disease

Ulcerative colitis (UC) and Crohn's disease (CD) are two forms of inflammatory bowel disease (IBD). They are two distinct diseases with different disease manifestations but with an overlapping clinical feature, chronic inflammation of the gastrointestinal tract (GI). IBD is a complex disease thought to be caused by a dysregulation in the mucosal immune response to commensal gut flora in a genetically susceptible host, which results in inflammation. Due to the early onset and the fluctuating disease course IBD has a substantial impact on a patients' quality of life.

1.1.1 Clinical presentation and treatment

IBD diagnosis is confirmed by clinical evaluation including relevant patient history, physical examination, laboratory testing, radiographic and/or endoscopic imaging and histology. Symptoms of IBD include diarrhoea, fever, abdominal pain, weight loss and rectal bleeding, with relapse and flare-ups often being part of the disease course. IBD is neither medically nor surgically curable; therapeutic approaches aim to induce and maintain symptomatic control, induce mucosal healing and improve quality of life. Therapeutic recommendations depend on the disease location, disease severity and disease-associated complications. Whilst a wide variety of effective therapeutics is available, surgery to remove damaged sections of the intestine is often needed (see section 1.1.1.3 for more detail). Although CD and UC are both forms of IBD, there are differences in disease manifestation and some treatments, although a successful differentiation between them cannot always be made.

1.1.1.1 Crohn's disease

In 1932 CD was first described by Burill Crohn, after whom the disease is named [1]. The disease is characterised by asymmetric, transmural and sometimes granulomatous inflammation which can affect any part of the gastrointestinal (GI) tract, creating patchy areas of inflammation [2,3]. The

terminal ileum is most commonly affected in CD and the earliest mucosal lesions tend to appear over the Peyer's patches [4]. CD is a phenotypically diverse disease with clinical manifestations, which include transmural inflammation, including strictures, lesions, penetrating fistulas, abscesses and, extra-intestinal manifestations (**Figure 1.1**) [3]. CD patients with a positive family history are more likely to have small bowel disease manifestations and demonstrate earlier disease onset [5]. Disease onset occurs in approximately 25% of cases during childhood; the remainder of cases occurs second and third decades of life. Age of onset does not appear to have an impact on increased disease development in IBD [6]. Therapeutic approaches attempt to induce and maintain clinical remission but in Crohn's disease this a real challenge and a large proportion of patients will relapse multiple times during their disease course. Disability caused by CD tends to be greater than by UC, with only 75% of patients being able to work in the year following diagnosis and 15% still unable to work after 5-10 years of disease [7].

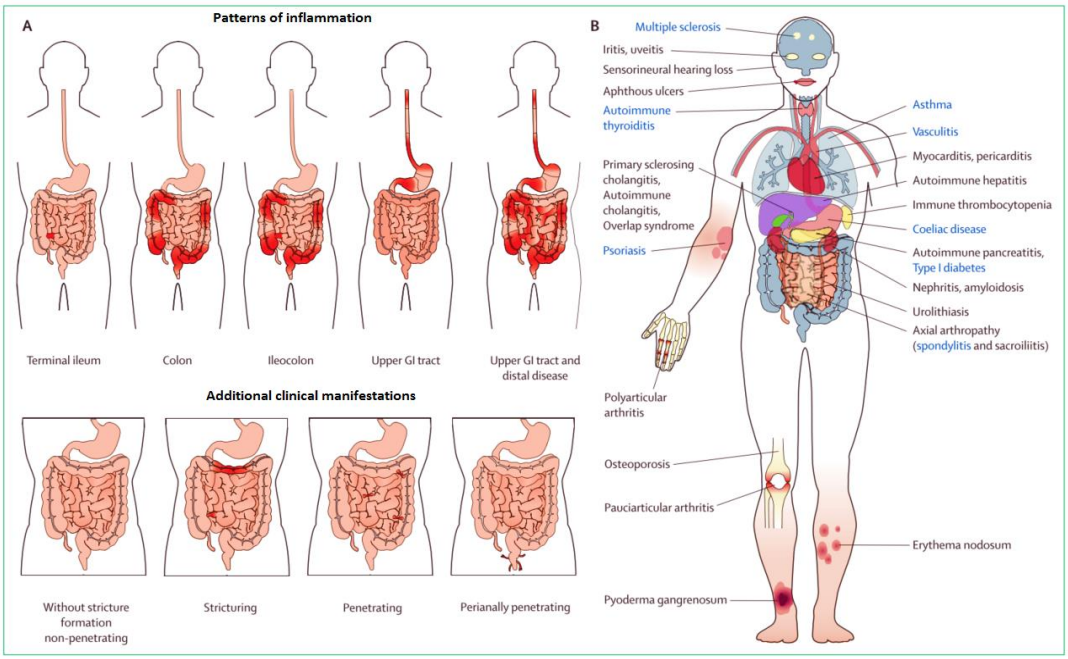


Figure 1.1 | Phenotype of Crohn's disease
(A) Various forms of CD disease manifestations including patterns of inflammation and additional manifestations. (B) Major extraintestinal manifestations and associated autoimmune disorders (blue). GI=gastrointestinal (Picture adapted from Baumgart *et al.*[3]).

1.1.1.2 Ulcerative colitis

Pinpointing the first observation and description of Ulcerative colitis is more difficult, although the British Physician Sir Samuel Wilks first refers to it by name in 1859 [8]. UC is a more homogenous condition with inflammation classically involving the rectum and extending proximally in a continuous manner, involving part of or the entire colon [9]. Disease extent is an essential factor in determining treatment in UC. UC can broadly be subdivided into proctitis (inflammation confined to the rectum), left-sided colitis (inflammation up to the splenic flexure), extensive colitis (inflammation up to the hepatic flexure), and pancolitis (involving the whole colon) (**Figure 1.2**) [10]. Patients with a more extensive form of UC occasionally display segmental inflammation, backwash ileitis or a caecal patch, which may lead to confusion with CD [11]. In contrast to CD, inflammation in UC only affects the surface mucosal layer of the bowel and does not extend to its full thickness. Onset for UC shows a bimodal pattern with the first and biggest peak of onset being between ages 15 and 30 years and a second smaller peak at age 50-70 years [12]. Therapies focus on inducing clinical remission and management of flare-ups.

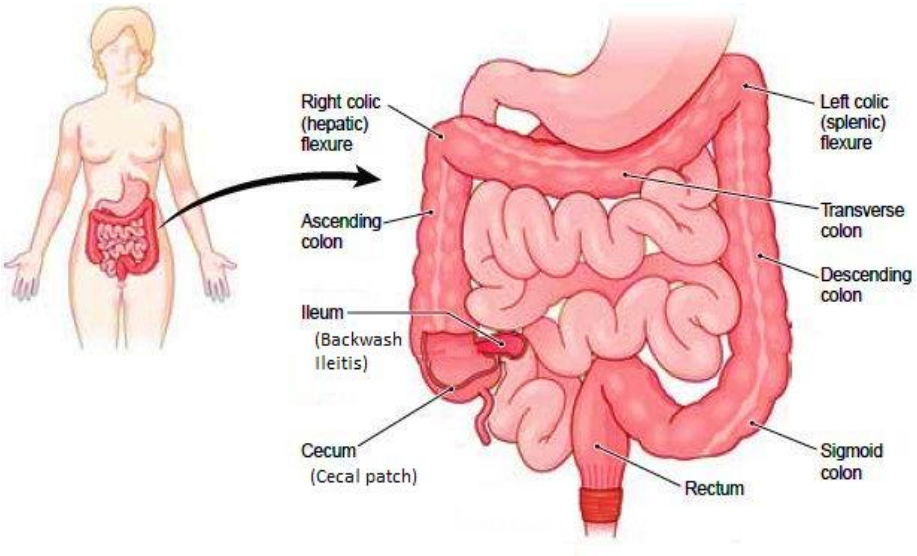


Figure 1.2 | Anatomy small and large intestine
Anatomy of the large intestine indication essential locations used when determining the extent of ulcerative colitis including the spelinc flexture and hepatic flexture as well as sides of segmental manifestations backwash Ileitis and cecal patch (picture adapted from http://encyclopedia.lubo-pitkobg.com/Stomach.html).

1.1.1.3 Treatment of IBD

There are three main aims in the treatment of IBD, achievement of remission, mucosal healing and maintenance i.e. prevention of disease flares. Clinical treatments to achieve this can be either medical or surgical in nature, with medical therapy being a rapidly evolving field in IBD over the last few years. Currently, the first-line of treatment in mild to moderate cases of CD exists of anti-inflammatory drugs such as corticosteroids or immunosuppressive drugs such as azathioprine, whereas the most common forms of treatment for UC are anti-inflammatory drugs such as aminosalicylates (ASAs) in mild to moderate cases and corticosteriods in moderate cases or for people not responding to ASAs. Immunosuppressive drugs such as azathioprine or cyclosporin are often used to maintain remission in UC if previous therapies have not been successful. Patients with more severe forms of IBD, or ones that have failed first-line treatment drugs, might require treatment with biologics. Biologics are a fast growing field of therapies with currently up to six agents approved for use in IBD [13]. The first biologics to the market were anti-tumour necrosis factor (anti-TNF) drugs of which three are now approved in Europe for induction and maintenance of remission: infliximab, adalimumab, and golimumab. In terms of maintenance, an infliximab study in CD showed a 12% loss of response per patient-year of treatment [14], whereas another study showed 55.7% sustained benefit after 5-years [15]. Within UC 30.6% of patients have been reported to stop infliximab within 3 years due to adverse events, lack of efficacy or other reasons [16].

More recently, vedolizumab an anti-integrin agent, was approved for treatment of IBD. Vedolizumab, an anti-alpha-4 beta-7 integrin antibody, decreases lymphocyte trafficking into the gut by preventing white blood cells from binding to the vascular endothelium through the interaction between alpha-4 beta-7 and MAdCAM, thereby limiting inflammation. It is a relatively slow acting drug, and therefore is often better for maintenance than induction [17]. Vedolizumab has shown higher efficacy in UC than CD with 47.1% vs 14.5% of patients showing clinical response at week 6 in clinical trials although real

life use suggests that its effectiveness in Crohn's disease may be better than its efficacy [18,19].

Lastly, an anti-p40 antibody, ustekinumab, which functions through the blocking of the p40 subunit of IL-12 and IL23 disrupting signalling though Th1 and Th17 pathways, has recently been approved for use in Crohn's disease. Ustekinumab has been shown to be affective for both induction (43%) and maintenance (50.5% of the 43% of patients showed successful induction) in CD patients over a year, and is effective in both biologic naïve patients and patients that have failed anti-TNF treatment [20]. Although, biologics offer great promise they might cause serious side effects including increased risk of cancers, congestive heart failure or serious infections [21]. Furthermore, biologics impose a considerable cost on the NHS [22] and other healthcare systems [23].

In addition to drug therapies, IBD patients often need surgery to treat the disease. Approximately, 70%-80% of CD patients require surgery during their lifetime [10], most commonly due to disease complications in the form of recurrent intestinal obstruction, strictures and perforations [7]. Surgery is not curative in CD; it is used to manage and minimise the impact of the disease. For UC, surgery rates at 10 years post diagnosis vary between 3% and 17%, with colectomy being needed either for severe UC or treatment-refractory UC in up to 27% of cases [24].

1.1.2 Epidemiology

The incidence and prevalence of IBD are highly variable between geographic regions and different ethnicities. The highest prevalence rates for both UC and CD are found in Europe (UC, 505 per 100,000 persons; CD, 322 per 100,000 persons) and Canada (UC, 248 per 100,000 persons; CD, 319 per 100,000 persons) [25]. Within Europe the highest prevalence and incidence is found within the United Kingdom [26,27] and Scandinavia [28-30] with the occurrence in Eastern Europe to remain rare [31], suggesting a North-West/South-East gradient in IBD incidence [32]. When evaluating the influence of ethnicity on IBD disease incidence, IBD is observed to be most common in Caucasians (324 per 100,000)

compared with Africans, Asians and Hispanics (239, 162 and 147 per 100,000), respectively [33]. Incidence rates for IBD have been increasing world-wide with 75% of CD and 60% of UC studies showing a statistically significant increase in incidence over a minimum 10-year period [25]. Specifically, developing nations with low incidence that have adopted an industrialised lifestyle see a major increase in IBD incidence, which suggests an important role of environmental factors in triggering the disease onset [25]. Nevertheless, twin studies have shown a definite genetic component to IBD with concordance rates in monozygotic (MZ) twins of 35-58% in CD and 6-13% in UC [34,35]. Additionally, having one or more affected first degree relatives still confers a greater risk than any known environmental factor [36,37], with the lifetime risk to offspring of two IBD affected parents exceeding 30% [38]. More recently Genome Wide Association Studies (GWAS) have been used to investigate genetic aetiology in IBD (see section 1.2.1). GWAS allows us to compare allele frequencies for each of more than 500,000 single nucleotide polymorphisms (SNPs) spanning the entire human genome in many thousands of disease cases and controls [39]. SNPs showing differences in allele frequencies between cases and controls will highlight regions of the genome that are associated with disease. To date 26% and 19% of heritability for CD and UC, respectively, has been explained by identified disease associated SNPs [40].

1.1.3 Environment

A clear genetic factor in the onset and occurrence of IBD has been established. Nevertheless, the incidence and prevalence of IBD has been increasing at a rate too rapid to be purely explained by a genetic effect, suggesting that other factors are involved in the occurrence of IBD. Environmental factors like smoking, appendectomy, diet and lifestyle have been implicated in IBD development.

1.1.3.1 Smoking

Smoking is one of the most studied environmental risk factors in IBD, demonstrating a paradoxical relationship between smoking and IBD. Through a discordant sib pair study Bridger *et al.* showed significant influence of smoking on the development of IBD; 21 out of 23 sibling pairs discordant for smoking at diagnosis showed non-smokers to develop UC and smokers developing CD ($p<0.0001$) [41]. A meta-analysis identified a direct association of smoking with a twofold increased risk in CD, whereas an association of the same magnitude was identified between never smokers or ex-smokers and UC, indicating a strong inverse association with UC [42]. The metabolite nicotine in cigarette smoke is most likely the main driver of the effects on IBD disease course, although studies examining the individual effects of nicotine and tobacco failed to recreate the same magnitude of association seen by smoking, suggesting that other components of tobacco smoke might be important [43,44]. Suggested ways of interaction include changes to the immune system, on both humoral and cellular level [45,46], altered cytokine production [47,48] and the body's increase in oxygen-free radicals production [49]. Additionally, changes to the intestinal mobility and permeability have been observed [50,51]. A gene expression study identified three genes significantly upregulated in CD smokers with active disease versus CD non-smokers [52], indicating a complex gene – environment interaction. The protective effect of smoking in the development of UC versus the causative effect in CD development highlights the difference in pathogenesis of the two diseases.

1.1.3.2 Appendectomy

Another factor which demonstrates a contradictory effect on CD and UC is appendectomy. Results are conflicting depending on age and reason for appendectomy, but it is reported that appendectomy prior to age 50 and for inflammatory reasons is associated with a protective effect for UC and a slight increase in risk for CD [53,54].

1.1.3.3 Diet and food antigens

Food antigens are the most common type of luminal antigen following bacterial antigens, making diet an important environmental factor to investigate. In addition, differences in diet might contribute to the geographical variance in IBD incidence. Case-control studies focusing on diet have numerous limitations. Nevertheless, associations between fibre, saturated fat and vitamin D intake have been reported. A 40% reduction in risk of CD was shown in women with long-term fibre intake [55], although it is difficult to establish if a reduced fibre intake increases risk or is a result of the disease. Human - as well as mouse - studies have indicated that high intake of saturated fat increases the inflammatory response and thus risk of IBD [56,57]. Furthermore, low levels of vitamin D are common in newly diagnosed IBD patients and are associated with hospitalisation and surgery in CD patient [58,59].

1.1.3.4 Lifestyle risk factors

Factors such as stress, exercise, sleep and hygiene can all be captured under the environmental risk factor lifestyle. Multiple studies have investigated the various environmental risk factors, although these studies come with limitations and data does not contribute to the treatment course. Of these, stress is the most convincing contributor to IBD, with studies suggesting an association between major life stressors or anxiety and IBD activity [60,61]. Gut inflammation can be influenced by stress through various neural components of the brain-gut axis, resulting in pro-inflammatory cytokine production, activation of macrophages, and changes in intestinal permeability and gut microbiota [62]. A higher risk of IBD development was noted in people with sedentary occupations, whereas, people with heavy labour jobs were at lower risk of IBD [63], implicating physical activity to have a beneficial effect. This was confirmed by a study reporting a 44% reduction in risk of CD in participants with high weekly activity versus inactive participants [64].

Sleep and hygiene are two factors that have proven more difficult to study. An association between disturbed sleep quality and active IBD active has been

reported, although this association might be bidirectional where higher disease activity causes poor sleep quality [65]. Studies investigating hygiene have mainly been focused on differences in exposure to certain organisms between developed countries and developing countries. It is hypothesised that the immune development is negatively affect by raising children in an extreme hygienic environment [66]. The use of chlorinated water and supermarket produce in developed countries reduces the exposure to organisms such as saprophytic *Mycobacteria, Lactobacilli* and *Helminths*. They are often harmless organisms, common in developing countries, found in mud, untreated water and fermenting vegetable matter. During exposure they interact with the innate immune system and induce T regulatory cells (Tregs), which in turn regulate anti-inflammatory cytokine IL10 production [67]. Absence of exposure to these organisms is thought to contribute to the inappropriate inflammation seen in IBD patients.

1.1.4 Gut microbiome

The gut microbiome is a collective name for the wide variety of bacterial species, viruses and fungi inhabiting the intestinal tract. The gut microbiota and their human host have had a symbiotic relationship and co-evolved for millennia, proving many benefits to its host. Nevertheless, microbes are foreign to the body and require tight regulation by the host immune system to maintain homeostasis. Disruption of this homeostasis can potentially lead to chronic inflammation. Microbiome composition varies depending on exposure to microbes via dietary intake and environmental factors as well as our genetic state, as shown by a large twin study [68]. By investigating the gut microbiome of 416 twin pairs, they identified microbial taxa who were highly heritable as well as microbial taxa which were influenced by diet and environment [68]. Furthermore, the gut microbiome has been shown to influence a host's metabolism and development of the immune system and its function [69]. An imbalance, or dysbiosis, of the microbiome can thus have broad effects on the general health of an individual. Normal gut flora can contain over 100 trillion

microbes, with the majority being bacteria. Healthy microbiota consists of nearly 1000 different bacterial species, with over 90% being from one of four phyla. Firmicutes (both Bacillus and Clostidiales) and Bacteroidetes phyla dominate with 30-60% and 20-50%, respectively, with Proteobacteria and Actinobacteria phyla accounting for a smaller proportion [70] (**Figure 1.3**).

1.1.4.1 Microbiome in IBD patients

When investigating the microbiota in IBD patients, an altered composition and reduced diversity of microbes is observed [71]. The most consistent observation in IBD patients is a reduction in Firmicutes and an increase in Proteobacteria (**Figure 1.3**) [72].

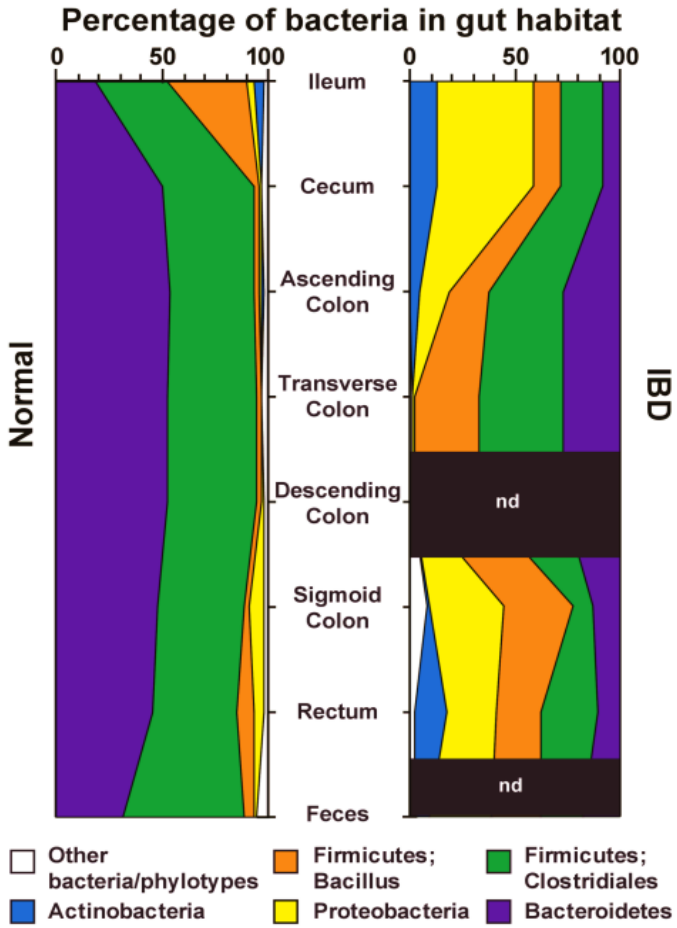


Figure 1.3 | Bacterial Phyla in the human microbiome
Relative abundance of predominant bacterial phylotypes in the human intestinal tract in relation to the location of the distal gut for healthy controls and IBD patients. (nd = not done) (picture adapted from Peterson *et al.* [72]).

Besides the altered diversity in IBD patients it has also been established that the microbiota in IBD patients – both with active disease and in remission - is less stable than in healthy people [73,74]. Beyond bacterial dysbiosis, non-bacterial microbes including viruses and fungi should be considered in IBD pathogenesis. Fungal dysbiosis has been reported in IBD patients compared to healthy controls, with CD patients gut environments maybe giving favour to fungi over bacteria [75]. It was suggested their might be disease-specific inter-kingdom alterations present in IBD gut microbiota [75]. Finally, there is the intestinal virome to consider; in depth analysis have shown that the enteric virome is abnormal in CD and UC patients [76]. Moreover, data suggests that changes in the virome may cause bacterial dysbiosis and contribute to intestinal inflammation [76]. Factors such as medication and inflammation are known to affect the microbiota, making it difficult to establish if the observed dysbiosis is caused by IBD or contributing to causing IBD.

1.1.5 Immune response in IBD

The intestinal tract is constantly exposed to a multitude of antigens, including food and bacterial antigens, requiring the host immune system to regulate an appropriate response. The layer of epithelial cells lining the gut wall is covered by a mucus layer secreted from goblet cells, which form the first line of defence by preventing direct interactions between luminal antigens and the host immune system. Intestinal inflammation in IBD is caused by a defect in epithelial barrier, combined with a dysfunctional response of the innate and adaptive immune system to commensal gut flora [77].

Epithelial cells are linked by tight junctions, preventing access of luminal antigens to the lamina propria. Damage to the epithelial barrier leads to increased permeability and enables the uptake of luminal antigens, triggering an immune response [78,79]. In addition to acting as a physical barrier, epithelial cells secrete various bactericidal agents including defensins. Reduced levels of human ß-defensin-1 (HBD-1) have been observed in both CD and UC, whereas,

human ß-defensin-2 (HBD-2) and human ß-defensin-3 (HBD-3) have shown lack of induction in CD but not UC [80].

The lamina propria, located beneath the epithelial cell layer, is densely populated with both innate and adaptive immune cells. The innate immune system includes immune cells such as dendritic cells (DCs) and macrophages, but also intestinal epithelial cells and myofibroblasts. Innate immunity provides rapid and effective inflammatory responses against microbial invasion. Innate immune cells express pattern recognition receptors (PRR), such as Toll-like receptors (TLR) and nucleotide binding domain (NOD) like receptors (NLR), with DCs containing the widest range of PRRs [81]. Within healthy intestinal mucosa TLR3 and TLR5 are primarily expressed, whereas, TLR2 and TLR4 expression is negligible [82]. In IBD an increased expression of TLR2 and TLR4 was reported [81]. Additionally, polymorphisms in NOD2, a NLR caspase recruitment domain, were highly associated with CD [83]. Activation through PRRs results in nuclear factor-$_k$B (NF-$_k$B) expression and production of pro-inflammatory cytokines [81]. It has been shown that patients with active UC and CD have higher levels of pro-inflammatory cytokines TNF-a, IL-6 and IL-8 due to a higher number of DCs present [84]. Furthermore, DC PRR activation triggers the adaptive immune system by initiating T cell differentiation.

The adaptive immune system, with T and B lymphocytes as the major cell types, is highly specific and creates long lasting immunity. The adaptive immune system is thought to be involved in the persistence of inflammation, with adaptive immune responses in IBD patients characterised by an imbalance of Tregs and effector T cells including CD4^{pos} T helper cells and CD8^{pos} cytotoxic T cells [85,86]. CD4^{pos} T helper cells have been shown to be important to CD as they represent the majority of the activated mononuclear cells that infiltrate the intestinal wall [87]. Where, CD8^{pos} cytotoxic T cells can be found in the mucosa in mouse models of IBD and several studies have suggested that autoreactive CD8^{pos} T cells may be involved in the initiation of the inflammatory response in IBD [88]. Furthermore, CD14^{pos} monocytes have been suggested to play an important role in intestinal inflammation. Macrophages (present in the lamina propria) expressing the CD14 (LPS) receptor are

markedly increased in tissues of patients with IBD [89]. They respond to microbial products such as LPS and interferon-γ (IFN- γ), and are extremely important in mucosal immunity [89].

Naïve T cells (Th0) in the adaptive immune system differentiate into one of three T cell sets: Th1, Th2 or Th17 cells. With Th1 cells being essential for elimination of intracellular pathogen, Th2 cells mediating allergic reactions and protecting against parasites, and Th17 contributing to the removal of extracellular fungi and bacteria [90,91]. Notably, studies have reported that CD is identified by an upregulation in Th1 cytokines (e.g. TNF-a, IFN- γ, IL-12) as well as Th-17 associated cytokines (e.g. IL-17A, IL-21 and IL-23), whereas UC is identified by increase of Th2 cytokines (e.g. IL-5 and IL-13) in inflamed mucosa [3,9,85,86]. Genetic studies investigating single nucleotide polymorphisms (SNPs) support the imbalance model by linking loci involved in Treg, Th1, Th2 and Th17 differentiation to IBD [92]. Changes in cytokine production, pro- and anti-inflammatory, have major downstream effects in causing and maintaining inflammation.

1.2 Genetics of IBD

1.2.1 Early linkage studies in IBD

Linkage mapping has proven to be a powerful tool for mapping highly penetrant disease loci, with more limited successes observed within complex diseases. Early linkage studies in IBD identified potential disease susceptibility loci on chromosomes 3,7,12 and 16 [93,94]. In 2001 the *NOD2* gene (nucleotide-binding oligomerization domain containing 2), also known as *CARD15,* was identified as the first CD susceptibility gene, using linkage analysis and positional cloning [83,95,96]. NOD2 to date, still accounts for the highest explained heritability of CD, with ~50% of CD patients carrying at least 1 mutation in NOD2 [97]. NOD2 is involved in recognising bacterial molecules within the intestine and stimulating an immune response, highlighting the importance of innate immunity within IBD. *NOD*2 variant Leu1007insC leads to a truncated protein, resulting in altered activation of NF-$_k$B following bacterial triggers [96].

Furthermore, susceptibility at several other loci was detected including chromosome 5q31 (IBD5), a region encoding for several immunoregulatory cytokines including IL-4, IL-5 and IL-13 which have been implicated in CD pathogenesis [98], chromosome 10q23 encoding discs large homologue 5 (DLG5) which is involved in the maintenance of epithelial integrity [99], and chromosome 6p (IBD3) containing the major histocompatibility complex (MHC) [100]. Although some progress was being made by linkage analysis in identifying regions of the genome which might contain IBD susceptibility genes, the introduction of genome-wide association studies (GWAS) in 2006 provided a leap forward in unravelling the complex genetics of CD and UC.

1.2.2 GWAS success in IBD

In very early GWAS, four new IBD susceptibility loci at genome-wide significance ($p<5x10^{-8}$) were identified, highlighting the power of the GWAS approach. New associations included *IL23R*, a protective variant involved in a signalling cascade which promotes inflammation and coordinates an adaptive immune response [101] and *ATG16L1* a protein-coding variant involved in the autophagosome pathway [102]. Additionally, an association was found with two locations devoid of any genes, referred to as gene deserts, on chromosome 5p13 and 10q21 [103,104]. The largest IBD GWAS in this first phase was done for Crohn's disease by the Wellcome Trust Case Control Consortium (WTCCC) [105], which identified 9 genome-wide significant associations, including the autophagy related gene *IRGM*. Further GWA studies increased the power of GWAS and this way identified many more IBD susceptibility loci. Currently, 27 GWA studies have been performed on CD and 21 on UC [106], with the major ones being by Jostins *et al.*, Liu *et al.*, Huang *et al*. and De Lange *et al.*

In 2012 a major IBD meta-analysis of 15 GWAS and follow-up study on cases and controls mainly of European origin was performed [92]. A total of 25,000 SNPs with at least nominal association in the meta-analysis were tested for association in and independent set of 14,763 CD, 10,920 UC cases and 15,9777 controls by genotyping on the Immunochip. This identified 163 IBD

susceptibility loci ($p \leq 5.0 \times 10^{-8}$) in at least one of the three analyses (Ulcerative colitis, Crohn's disease, IBD). Of these loci, 110 were associated with IBD (i.e. both UC and CD) whereas 30 loci were classified as Crohn's-disease-specific and 23 as ulcerative-colitis-specific [92]. Furthermore, 38 additional IBD susceptibility loci have been identified by expanding the 2012 meta-analysis by Jostins *et al*. with 11,535 individuals of European descent and 9,846 individuals of non-European descent [107], in the first trans-ancestry GWAS (**Figure 1. 4**). De Lange *et al.* identified a further 25 IBD susceptibility loci by expanding the GWAS analysis by 25,305 individuals, bringing the total to 224 IBD susceptibility loci [108].

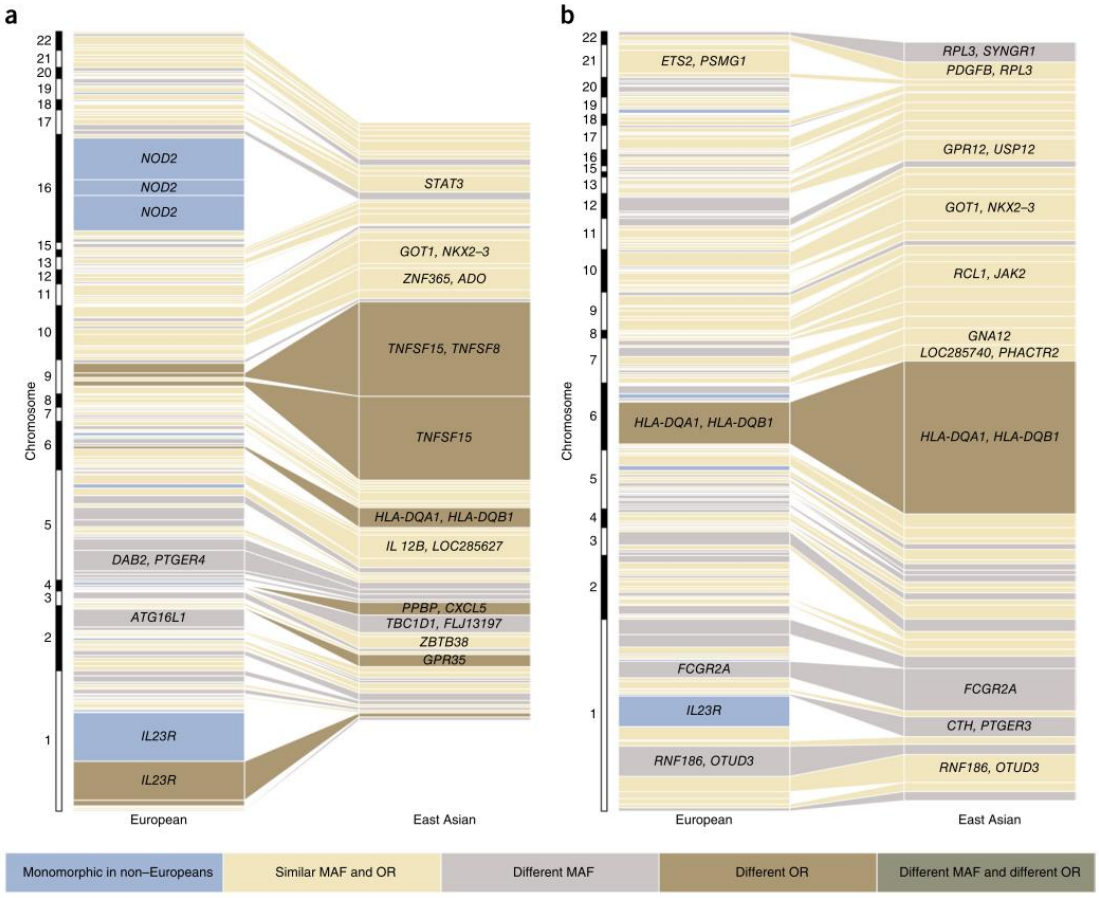


Figure 1.4 | Variance explained by the 200 IBD loci
Comparison of variance for Crohn's disease **(a)** and ulcerative colitis **(b)** between East Asians and Europeans. Each bar represents an independent disease associated locus. The width of the bar is proportional to the explained variance of that locus in East Asians or Europeans. Associations between both ancestries are represented by connecting lines, and the colour of each box indicates if any difference in variance is due to difference in allele frequency. MAF=Minor Allele Frequency, OR=odds ratio (picture adapted from Lui *et al.* [107]).

In an attempt to prioritise genes for causality within the ~220 identified IBD susceptibility loci, various analysis were performed including GRAIL (Gene Relationships Across Implicated Loci), DAPPLE (Disease Association Protein-Protein Link Evaluator), and eQTL (expression Quantitative Trait Loci) [107,109,110]. Approximately, 400 genes were prioritised for causality, with associated pathways underlying IBD susceptibility. Suggested pathways are involved in innate immunity, JAK/STAT signalling, cytokine production, lymphocyte activation and epithelial barrier function [92]. Prioritised genes within the 64 newly identified loci enforce previous finding, with *ATG4B* playing a key role in autophagy, *OSMR* modulating the epithelial barrier function, *SLAMF8* disrupting the migration and inflammatory response of myeloid cells, *RORC* regulating Th17 cells, and various other prioritised genes showing involvement with both $CD4^{pos}$ and $CD8^{pos}$ T cell responses [107,108].

Although, above mentioned methods have made some progress in prioritising causal genes, efforts are being made to identify causal genes and variants within the IBD loci via fine-mapping. Fine mapping or association mapping statistically identifies minor differences in strength of association between variants with high levels of correlation to infer which is likely to be causal [111]. Ninety-seven regions previously associated with IBD and containing at least one associated variant were chosen for a fine mapping analysis by the International IBD Genetics Consortium using dense sets of SNPs across these regions on the Immunochip [112]. Within 3 out of 97 regions no consistent credible association could be identified, whereas 139 independent associations were defined across the remaining 94 regions. Bayesian analysis was used to estimate the posterior probability of causality for associated SNPs in these regions. This way 45 out of 139 single associations were refined to a single causal variant with >50% probability, of which 18 associations had a probability of >95% to be the causal variant. Of these 45 most credible associations, 13 caused protein-coding changes, 3 disrupted a transcription factor (FT) binding site, 10 fell within tissue specific epigenetic markers and 2 showed co-localisation with a significant *cis*-eQTL. The remaining 21 single causal variants mapped to non-coding variants that are not located within

known motifs, annotated elements or are involved in known eQTLs [112]. While physical proximity does not guarantee functional relevance, fine-mapping has converted 94 IBD susceptibility loci into statistical convincing causal variants. Fine mapping can provide a powerful tool in guiding further experiments to investigate disease mechanism.

Despite these efforts fine mapping studies have shown that the majority of IBD associated SNPs are correlated with non-coding variants that may perturb regulation of gene expression instead of directly altering gene structure and function [92,112]. This is strengthened by the fine mapping findings by Farh *et al.*, showing that ~90% of IBD causal variants are non-coding [113]. Therefore, it seems reasonable to begin to shift focus from GWA studies to gene expression studies, in order to attempt to understand how the majority of common susceptibility loci may influence IBD pathogenesis.

1.2.3 Functional mapping of GWAS loci

1.2.3.1 Gene expression in IBD

To date, GWAS findings have facilitated a number of functional/gene expression studies in IBD but these were generally focused on a limited number of target genes. One such study looked at known CD risk variants on chromosome 5q33.1 located upstream of the autophagy gene *IRGM* (Immunity related GTPase related family, M). They investigated correlations between the risk haplotypes and gene expression, demonstrating that the CD risk haplotype was associated with a significant decrease in *IRGM* expression ($p<10^{-12}$) in untransformed lymphocytes from CD patients [114,115]. Various other studies have investigated the biochemical mechanism by which IL23R variants might provide protection against IBD, with the most recent being by Sivanesan *et al.* [116]. IL23R (interleukin 23 receptor), involved in initiating the differentiation of helper T cells (Th17), was one of the first CD susceptibility genes to be confirmed by multiple GWAS [101]. Investigation of multiple potential causal variants in IL23R, showed a reduction in IL23 mediated IL23R activation and subsequently a reduction in inflammatory response. It was established that the

observed reduction in IL23R signalling was due to lower levels of cell surface receptor expression [116]. Ellinghaus *et al*. showed that a significant reduction in expression of *PRDM1* (PR domain-containing 1) was associated with the CD risk C allele in both ileal biopsy specimens and peripheral blood mononuclear cells (PBMCs) (combined $P=1.6x10^{-8}$). A reduction in *PRDM1*, which encodes for a master transcriptional regulator in B and T cell, resulted in increased $CD4^{pos}$ and $CD8^{pos}$ proliferation, INF-γ secretion and upregulation of activation markers upon stimulation [117].

1.2.3.2 Differential expression analysis in IBD

The majority of differential expression analysis is performed at a small scale i.e. within single genes or groups of functionally similar genes suspected in IBD pathogenesis as part of validation and functional studies. Although, Granlund *et al.* performed a genome wide microarray-based expression study investigated gene expression in both inflamed and un-inflamed mucosa from 63 patients with UC, CD and 20 controls [118]. They identified 3 and 0 differentially expressed genes in the uninflamed CD and UC vs control analysis, respectively. Within the inflamed tissue they identified differential expression within 8,539 genes, with many antimicrobial peptides to be upregulated in inflamed tissue. Additionally, they reported altered gene expression in multiple IBD-associated genes between IBD cases and controls, with limited differences in gene expression pattern between UC and CD. Genes identified to exhibit differential expression within IBD cases and controls showed involvement in immunological and defence-related roles [118]. Furthermore, Peloquin *et al.* investigated differential expression of 678 IBD disease-associated genes, often not covered by microarray platforms, within 1,100 inflamed and non-inflamed mucosal biopsies [119]. They identified 431 DE genes in CD and 439 DE genes in UC colonic tissue vs controls, with 88% overlap between the CD and UC DE genes [119]. Two main findings were the observed downregulation of *VDR* (Vitamin D Receptor) and *SLC22A5* (Soluble Carrier Family 22 Member 5) in both UC and CD cases vs controls [119]. VDR has been reported to mediate microbe-host interaction through regulation of autophagy [120] and SLC family

members have previously been reported to be associated with IBD [121]. Furthermore, one of the genes reported to be differentially expressed between UC and CD is *INF4* (IFN regulatory factor 4). IRF4 is a master transcription factor for Th17 cells [122], a cell type shown to be important in inflammatory responses in CD [92]. Both above mentioned studies show that differential expression analysis of IBD cases vs controls in either inflamed or non-inflamed tissue can contribute to prioritise candidate genes and to inform future functional studies.

1.2.3.3 eQTL studies

GWAS and fine mapping studies have established that the majority of IBD associated genetic variants are located in non-coding regions of the genome [112,113]. This indicates that the majority of identified IBD susceptibility SNPs perturb regulation of gene expression in some way. Expression quantitative trait loci or eQTL studies are used to identify potential causal genes in disease by correlating genetic variation with alterations in gene expression [123]. eQTL can exist either in *cis* i.e. within 1MB on either side of the genetic variant, or in *trans* i.e. further upstream/downstream or on a different chromosome. So far it is suggested that the majority of eQTLs are in *cis*, although it is hypothesised that this might be due to lack of power and low sample size [124]. Further increased sample size can address the current enrichment of *cis* vs *trans* eQTLs reflects true biological processes or insufficient power to detect *trans* eQTLs.

Early eQTL studies, using lymphoblastoid cell-lines, have been very valuable in identifying multiple eQTL loci throughout the human genome [124-126]. The Genotype-Tissue Expression Project (GTEx) characterized eQTLs across 44 tissues in approximately 449 individuals [127], establishing a major public resource database showing cell-type gene expression signatures. Recent studies have shown the importance of cell-type specific gene expression signatures with a study comparing eQTLs from 43 different autopsy-derived tissues including PBMCs, detecting only a 50% overlap in eQTLs within 9 tissues sampled in >80 donors [128]. Another study compared *cis* eQTLs from three

different cell types from the same individuals; fibroblasts, LCL's and primary T cells, reported a 69-80% cell type specificity in eQTLs [129].

Several studies investigating the presence of eQTL correlated with IBD associated variants have now been performed [130-132]. An early eQTL study by Kabakchiev *et al*. investigated eQTL within human small intestine [130]. They identified more than 15,000 statistical significant *cis*- and *trans*-eQTLs, of which 30% to 40% have previously been identified as eQTLs in various other tissues [130]. In addition to identifying eQTLs in ileal tissue, they investigated the presence of eQTLs specific to IBD risk loci. 155 IBD-associated SNPs were associated with altered gene expression levels of genes within a 50 kb window of each SNP, 27 significant *cis*-acting eQTLs were identified [130]. A study by Repnik *et al.* investigated 9,563 eQTL correlations with 402 IBD associated SNPs located within 208 candidate loci within intestinal tissue and PBMCs. They were able to confirm multiple previously suggested eQTLs at loci which included *SLC22A5*, *ECM1* and *PUS10*. Additionally, they identified a novel eQTL correlation with *ECM1* on chromosome 1q21 [131]. Furthermore, Singh *et al*. [132] performed an eQTL study in 39 IBD patients and 33 controls on tissue collected from their terminal ileum and four colonic locations. Approximately 1,871 independent *cis* eQTLs were found throughout the colon at a false discovery rate (FDR) of 5%, with the majority identified in rectal mucosa. 27% of eQTLs found in the rectal dataset were novel when compared to 7 datasets from various other tissues, the rectal dataset did show enrichment for genes known to be expressed in the colon [132]. When investigating eQTLs within IBD associated loci only 11 eQTLs were found, 6 of which (*ERAP2*, *SFMBT1*, *FUT2*, *ADCY3*, *INPP5E*, and *UBE2L3*) have been previously identified in the Ilieum [130], four (*CPEB4*, *IRF5*, *ATG16L1*, and *TSPAN14*) were previously identified in other tissues, and one novel eQTL, *STX4*, was identified [132]. *STX4* (Syntaxin 4) is involved in regulation of secretion from various immune cells [133]. More recently, a study by De Lange *et al.* [108] employed eQTL analysis to investigate variant-gene associations within 25 newly identified IBD risk loci. Associations between nearby genes and the IBD index SNPs (or SNPs with linkage disequilibrium (LD) of $r^2 \geq 0.8$) were investigated within 12 eQTL databases.

Overall, 19 significant eQTLs were identified within 10 out of 25 IBD risk loci [108]. Furthermore, they investigated associations within four identified integrin genes (*ITGA4, ITGAV, ITGB8* and *ITGAL*) and their binding partner ICAM1 [108]. Three out of the five associations were observed to be driven by the same variants as LPS monocyte-specific stimulus response eQTLs. This suggests upregulation of pro-inflammatory cell surface markers are a potential mechanism of action [108].

eQTL studies and their efforts to identify causal genes within complex diseases are the next big thing in genetic studies. The main things to consider going forward in eQTL studies for mapping IBD causal genes are: disease specific tissue, heterogeneity of cell types in tissue and statistical power. Pinch biopsies used to identify eQTLs within intestinal tissue are not homogenous and contain various cell-types including epithelial cells, stromal cells, and various immune cells. Gene expression signatures that are identified therefore fail to provide a uniform picture. Primary sorted cells of heterogeneous tissues might give a clearer picture of eQTLs in colon tissue but experimentally this is not as straightforward.

1.2.4 Biomarkers in IBD

Biomarkers are measurable indicators of a biological state and are commonly used within a clinical setting as indicator of disease. The use of biomarkers to aid diagnosis of IBD and the differentiation between UC and CD has been increasing over time, although no single biomarker has proven strong enough to be used by itself. C-reactive protein (CRP) and erythrocyte sedimentation rate (ESR), are two common biomarkers indicative of general inflammation which are used in IBD as well as other inflammatory diseases [134]. CRP is a component of the innate immune system and is produced by hepatocytes in response to specific pro-inflammatory cytokines [135]. CRP base levels vary between people but higher levels of CRP most likely indicate the presence of an inflammatory response. Increased levels of CRP are more often seen in CD than UC at the time of diagnosis, and within known IBD patients an increase of

CRP levels often correlates with active disease [136]. ESR evaluates the rate at which erythrocytes fall through plasma, which depends largely on the fibrinogen concentration within plasma.

pANCA (perinuclear antineutrophil cytoplasmic antibody) and ASCA (anti-*Saccharomyces cerevisiae* antibody) are antibody biomarkers which can help to distinguish between UC and CD. A study demonstrated that 64% of UC patients were pANCA positive and ASCA negative [137], whereas CD patients were identified as pANCA negative and ASCA positive with 93% specificity [138]. Unfortunately, pANCA/ASCA has a low sensitivity, 55% in CD [138], preventing it from routine clinical use.

In IBD patients, stool based biomarkers, identifying gastrointestinal inflammation, have proven very valuable. An intestinal inflammatory biomarker routinely used in clinical practice is calprotectin. Faecal calprotectin is a zinc and calcium protein derived from neutrophils, monocytes and activated macrophages [139]. During the inflammatory process calprotectin is released by mucosal epithelial cells and through degranulation of neutrophils within the intestine [140]. Calprotectin levels are a relatively reliable indicator of inflammation in IBD; nevertheless, it will always be used in combination with endoscopy and histology to make the diagnosis [141].

Due to the relative lack of specificity of currently available biomarkers to differentiate between IBD subtypes or to predict prognosis and relapse rates, the development of disease specific biomarkers is of great clinical interest. Various research studies into new antibody or protein biomarkers for IBD are underway and interestingly, the first use of genetic markers as biomarkers has been reported (see section 1.2.5). Successful development of novel biomarkers to aid better prognosis and prediction of relapse could have a huge impact on patient treatment and wellbeing.

1.2.5 Transcription signatures as biomarkers in IBD

In order to facilitate the need for disease specific biomarkers in IBD, the use of gene expression profiles to detect novel transcription signatures with biomarker potential has been investigated. Several studies have generated expression profiles to aid biomarker development, but so far with limited success. This is most likely due to the heterogeneous state of the samples used within these studies, e.g. peripheral blood mononuclear cells (PBMCs) or mucosal biopsies. Transcriptional variation across multiple genes will predominantly reflect differences at a cellular level; thus separation of various cell subsets will allow these differences to be more easily detected. Another factor suggested to affect the success of identification of transcriptional signatures as biomarkers is the activation state of the sample used. Inflamed tissue samples as well as *in vitro* stimulated PBMCs show stronger expression signals of immunological and/or pro-inflammatory IBD-associated genes, suggesting that gene expression studies in inflammatory diseases require activation for signals to protrude from background.

Von Stein *et al.* in 2008 identified a 7 gene panel differentially expressed between CD and UC patients with active disease [142]. Their small scale follow-up study reported 90% success in diagnosing UC, IBD and non-IBD patients based on those 7 genes [142]. More recently, the gene panel was tested in inflamed colonic tissue from 78 difficult to diagnose patients, leading to a change in primary diagnosis for a significant number of patients [143]. Although, the diagnosis by biomarker remained different from the primary diagnosis in 6% of UC and 5% of CD patients. Furthermore, Lee *et al*. using gene expression microarrays, has identified a transcriptional signature based on the expression profile of 14 key genes in separated $CD8^{pos}$ T cells that was significantly associated with altered prognosis. The cohort presenting with this signature had higher incidence of experiencing treatment-refractory, relapsing, or chronically active disease in both CD and UC [144]. Furthermore, a milder course of CD has been associated with a noncoding polymorphism in the *FOXO3A* gene (combined $p = 2.1x10^{-8}$) [145]. Monocytes containing the polymorphism were

observed to produced twice as much FOXO3A upon stimulation. More FOXO3A in monocytes reduces production of pro-inflammatory cytokines and increases production of anti-inflammatory cytokines, thus, resulting in a milder course of CD [145].

The outcomes of these studies have shown that gene expression profiling may have great potential for the development of biomarkers which can aid in prognosis, diagnosis and relapse within IBD.

1.3 Aims

This PhD consists of two separate projects, the aims for which are outlined below.

1.3.1 Quantitative and qualitative analysis of the transcriptome in colon

The aim of this project was to increase the knowledge of the pathogenesis of Inflammatory Bowel Disease (IBD). Firstly, a qualitative analysis was performed to examine expression levels of genes and non-coding RNAs at the known IBD susceptibility loci in biologically relevant intestinal tissue from affected patients and controls. Secondly, genes exhibiting differential expression between IBD cases and controls were investigated within the colon and biological pathways affected by the differentially expressed genes were examined. Thirdly, a genome wide expression quantitative trait (eQTL) analysis were carried out in colonic tissue to look for association of IBD risk SNPs with altered gene expression. Finally, expression deconvolution was performed to assess the effects of colonic biopsy cell composition on gene expression.

1.3.2 Investigation of transcriptional biomarkers in prediction of relapse

The aim of this study was to identify a panel of new and effective biomarkers for the prediction of likely relapse in patients suffering from Crohn's disease. Cell type specific transcriptional profiles were generated for unstimulated and stimulated $CD4^{pos}$ T helper cells, $CD8^{pos}$ cytotoxic T cells and $CD14^{pos}$ monocytes using Illumina HT-12 expression bead arrays. A comparative analysis of the generated transcription profiles will be performed for patients relapsed vs non-relapsed for each cell type and for both unstimulated and stimulated cells.

2. Materials and Methods

2.1 Materials

2.1.1 Reagents

Reagent	Manufacturer
4'-6-diamidino-2-phenylindole (DAPI)	AnaSpec
CD3/CD28 T-activator beads	Thermo Fisher
Chloroform	VWR
Collagenase Ia	Sigma-Aldrich
Dimethyl sulfoxide (DMSO)	Sigma-Aldrich
DNase I	Roch
Ethanol	Fisher Scientific
Ethylenediaminetetraacetic acid (EDTA)	Ambion
FcR block	Miltenyi
Fetal bovine serum (FBS)	Life Technology
Gentamicin	Sigma-Aldrich
Hank's Balanced Salt Soln (HBSS) Medium w/o Mg^{+}/Ca^{+}	Fisher Scientific
Isoamylalcohol	VWR
Lymphocyte separation media	MP chemicals
Lypopolysaccaride (LPS)	MP chemicals
Magnesium Chloride ($MgCl_{2)}$	BioLine
Penstrep	Life Technology
Phosphate-buffered Saline (PBS)	Sigma-Aldrich
Proteinase K	Fisher Scientific
QIAzol lysis reagent	Qiagen
RNAlater®	Ambion
RNase free DNaseI	Qiagen
RNeasy mini Kit	Qiagen
RNeasy miRNA micro Kit	Qiagen
Roswell Park Memorial Institute (RPMI) 1640 Medium w/o Mg^{+}/Ca^{+}	Sigma-Aldrich
Sodium Chloride (NaCl)	Sigma-Aldrich
Sucrose	Sigma-Aldrich
Tris	Ambion
Triton x-100	Sigma-Aldrich
Trypan blue dye	Thermo Fisher

2.1.2 Solutions and media

10x TRIS buffer
109.5 g sucrose
10 ml 1M Tris (pH 7.5)
5 ml 1M MgCl2
10 ml Triton x-100
per 1 litre MilliQ H20

1x SET buffer
5.48 g NaCl,
10ml 1 M Tris (pH7.5)
2ml 0.5 M EDTA (pH 8.0)
per 1 litre MilliQ H20

Tris-EDTA buffer
10ml 1 M Tris HCL (pH 7.5)
2ml 0.5 EDTA (pH 8.0)
per 1 litre MilliQ H20

Collection media
HBSS w/o Mg^{+}/Ca^{+}
1% Penstrep
1.2 mg/ml Gentamicin
1 mM EDTA

Digest media
RPMI1640 w/o Mg^{+}/Ca^{+}
10%FBS + 1% Penstrep
1.2 mg/ml Gentamicin
1 mg/ml filtered collagenase Ia
10 U/ml DNase I

Culture media
RPMI1640 w/o Mg^{+}/Ca^{+}
10% FBS
1% Penstrep
1.2 mg/ml Gentamicin

FACS buffer
1x PBS
10% FBS
2 mM EDTA

PBMC media
RPMI1640
2% FBS

Cell sorting solution
PBS
1% FBS
1mM EDTA

2.1.3 Antibodies and primers

Table 2.1| antibodies used for cell sorting and biopsy immophenotyping

Antigen	Conjugate	Isotype	Supplier	Cat. No.
CD3	PerCP/Cy5.5	Mouse IgG2a, $_k$	BioLegend	317336
CD4	APC	Mouse IgG2b, $_k$	BioLegend	317416
CD8	FITC	Mouse IgG2b, $_k$	BioLegend	344704
CD14	PE	Mouse IgG1, $_k$	BioLegend	325606
CD45	APC/Cy7	Mouse IgG1, $_k$	BioLegend	304014
CD66b	PerCP/Cy5.5	Mouse IgM, $_k$	BioLegend	305107
CD68	PE-Cy7	Mouse IgG2b, $_k$	BioLegend	333815
CD90	APC	Mouse IgG1, $_k$	BioLegend	328113
CD326	PE	Mouse IgG2b, $_k$	BioLegend	324205
Mouse IgG1, $_k$	APC	-	Biolegend	400121
Mouse IgG1, $_k$	APC/Cy7	-	Biolegend	400127

Table 2.2 |Primers used for assessing TNF expression

Gene	Conjugate	Supplier	Cat. No.
RP18S	FAM	ThermoFisher scientific	Hs99999901_s1
TNFα	FAM	ThermoFisher scientific	Hs00174128_m1

2.2 Methods Project 1: Quantitative and qualitative analysis of the transcriptome in the colon

2.2.1 Power Calculation

Standardised methods to calculate power and samples size in differential expression analysis based on negative binomial distribution using RNA sequencing data are currently lacking. Such power calculations are dependent on multiple factors including the sequencing depth, the number of biological replicates and the level of expression of the genes e.g. lower expressing genes will require larger sample sizes to reach 80% power than highly expressed genes. Multiple studies have attempted to address this issue and performed power and sample calculation for RNA sequencing differential expression analysis [146,147]. Ching *et al*. established that a sequencing depth of > 20 million reads and ~25 samples per condition were needed to reach 80% power using EdgeR (differential expression analysis tool) [146]. 24 controls, 28 UC and 76 CD samples were sequenced, a differential expression analysis on these cohorts should provide us with 80% power to detect >0.8 logFold changes. It should be considered that these papers only present estimates and that factors such as sample heterogeneity or disease state can affect the power.

Power calculation for the eQTL analysis was performed, using Genetic Power Calculator (http://statgen.iop.kcl.ac.uk/gpc/). It was calculated that analysis of gene expression in 105 samples will provide 80% power to detect an effect if the relevant SNP accounts for just 15% of trait variance (gene expression) and 93% power to detect an effect of 20%. 24 controls, 28 UC and 76 CD samples were sequenced, giving and overall samples size of 128 patients for the eQTL analysis.

The presence of a transcriptional signature predictive of relapse will be assessed through differential expression analysis between the remission and relapse patient samples, with the non-relapsed samples acting as an internal control

for natural transcriptional changes over time. Our overall patient cohort existed of 50 patient's samples, 33% of which relapsed. Taking into consideration the three different patient cohorts these patients were collected from; post-surgery patients, patients withdrawn from anti-TNF therapies and routine gastroenterology patients as well as overall relapse rates, this study is not sufficiently powered and should be considered a pilot study.

2.2.2 Sample collection

All 171 intestinal biopsy and peripheral blood samples were collected during routine colonoscopies at Guy's and St Thomas' Hospital after informed consent was given. Samples of 4-8 pinch biopsies from un-inflamed large intestine were collected from random sites throughout the transverse and descending colon and stored in 500µl RNAlater® (Ambion) at -20 °C until RNA extraction. 2.5 ml of blood was collected in Paxgene tubes (QIAGEN GmbH) for RNA extraction and 10 ml of blood was collected in EDTA tubes (BD Biosciences) for DNA extraction. Both were stored at -20 °C. Samples were collected over a two-year period by myself (including taking consent from patients and attending their colonoscopy), with help from Ariella Amar and the medical staff of the endoscopy units. The sample cohort included samples from diagnosed CD (108) and UC (32) patients, as well as a control/non-IBD group (31) having a colonoscopies for various medical indications (**Table 2. 3**). The imbalance in patient sample numbers per cohort was partly due to the original study design; previously the focus lay on CD patients only. Furthermore, the availability of patients played a role; there is a higher percentage of CD patients *versus* UC and non-IBD patients having colonoscopies. An almost equal spread between males and females was observed; the CD cohort contained slightly more males (54% vs 46%), where the UC and control cohorts included slightly more females (46% and 42% *versus* 54% and 58%) (**Table 2. 3**). The mean age within the cohorts was approximately 40 years of age, with UC patients being slightly older at mean 44.1 years (**Table 2. 3**). Treatment regimens, from patients at the time of the biopsy being taken, were not recorded.

Table 2. 3 | Patient demographics

	CD patients	UC patients	Controls (non-IBD)
Mean age	38.5	44.1	41.4
# Female	35	15	14
percentage	46%	54%	58%
# Male	41	13	10
percentage	54%	46%	42%

2.2.3 RNA preparation

RNA was purified from colonic pinch biopsies using an RNeasy Plus Universal Mini Kit according to manufacturer's instructions (QIAGEN Gmbh). Briefly, the samples were homogenised using a micro-tube pestle following addition of 900 µl QIAzol lysis reagent. Multiple wash steps were performed on column using RW1 and RPE buffer. Genomic DNA was removed using RNase free DNaseI (Qiagen). Lastly, 60 µl RNase-free water was added during two elution steps, to elute the RNA.

2.2.4 Quantifying RNA

The Qubit 2.0 Fluorometer (Invitrogen) was employed to quantify RNA yield. Either using the Qubit RNA Broad Range (BR) assay kit or Qubit RNA High Sensitivity (HS) assay kit, according to manufactures protocol.

2.2.5 Quality control RNA

The Agilent 2100 Bioanalyser, RNA 6000 Nano chip, was used to test RNA quality (RIN), according to manufactures protocol.

2.2.6 DNA preparation from whole blood

DNA was extracted from 10 ml of peripheral blood collected in EDTA tubes (BD biosciences) by means of sodium chloride (NaCl) precipitation followed by phenol: chloroform purification. 45 ml lysis buffer was added to whole blood prior to 20 min centrifugation at 2000 rpm and 4˚C. Samples were then

incubated overnight at 37˚C following the addition of 4.5 ml 1x SET, 250 μl 10% SDS and 100 μl proteinase K to achieve nuclear lysis and inhibition of nucleases. 2.5 ml saturated NaCl solution and an equal volume of isoamylalcohol: chloroform (1:24) solution was added prior to a 30-60 min incubation at RT under rolling condition followed by a 10 min centrifugation at 2,000 rpm, 21˚C. The supernatant was transferred and 2 volumes of 100% ice-cold ethanol were added. The precipitated DNA was washed in 70% ethanol and re-suspended in Tris-EDTA buffer, before storage at -20˚C. Qubit® 2.0 Fluorometer (Invitrogen) was used (under standard operating procedures) to determine the DNA concentration and quality.

2.2.7 Genome-wide SNP genotyping

Paired blood samples were collected in EDTA tubes for DNA extraction from 122 out of 127 patients and controls. The 122 DNA samples were processed for genotyping on either the Infinium Human Core Exome array (Illumina) (n=71) or Infinium Human Core (n=15) bead chip (Illumina) by the GSTT/KCL BRC Genomics Centre. The remaining 36 samples were genotyped by the Sanger Centre using the Human Core Exome bead chip via pre-existing collaboration with the UKIBD consortium. The Human Core Exome chip contains 551,839 SNPs (largely overlap with Human Core Chip SNPs) and the Human Core Chip contains 306,670 SNPs. When reviewing the genotype results, one sample was observed to have failed to generate any genotype data and was subsequently removed from the analysis. Genome studio was employed for genotype calling and to export genotype data to PLINK prior to use.

2.2.8 RNA sequencing library preparation

2.2.8.1 Ribosomal RNA depletion

Both human and bacterial ribosomal RNA (rRNA) was removed using the Ribo-Zero Magnetic Gold Epidemiology Kit (Epicentre, Illumina), according to manufacturer's instructions. 1 μg RNA (RIN $\geq$ 6) starting material was used at a 40 ng/μl concentration. Following, rRNA removal the RNA samples were

purified using RNeasy MinElute Kit (Qiagen), by the manufacturers protocol. The Agilent 2100 Bioanalyser, RNA 6000 Pico chip, was used to confirm the efficiency of the rRNA removal and the purity of the remaining RNA.

2.2.8.2 Epicentre ScriptSeq V2 RNA-seq library preparation

Following rRNA removal whole RNA libraries were prepared using ScriptSeq v2 RNA-seq Library Preparation kit (Epicentre, Illumina), according to the manufacturer's instructions. Briefly, the RNA is fragmented and primers were annealed using fragmentation solution and cDNA synthesis primers. Additionally, a 1:1000 dilution of External RNA Control Consortium (ERCC) Synthetic Spiking Standards (LifeTechnologies) was added prior to an 85˚C incubation for 5min. The cDNA was synthesised and terminal-tagged by adding cDNA Synthesis Premix, DTT and ScarScript Reverse Transcriptase prior to thermocycler incubation of 5min at 25˚C followed by 20 mins at 42˚C. The reaction was then cooled to 37˚C and Finishing solution added. The thermocycler programme was continued with a 10 min incubation at 37˚C and 95˚C for 3min. The reaction was then cooled to 25˚C and Terminal Tagging Premix and DNA polymerase was added. Finally, the reaction was terminated by incubating at 25˚C for 25 min, 95˚C for 3min and then cooling to 4˚C. The cDNA was purified using MinElute Purification Kit (Qiagen), according to manufacturer's protocol. The di-tagged cDNA was amplified and barcoded by addition of FailSafe PCR PreMix, FailSafe PCR Enzyme, Forward primer and the appropriate Set 1 Index reverse primer (Epicentre, Illumina), PCR was performed under the following conditions: 1 min at 95˚C, 15 cycles of 30 sec at 95˚C, 30 sec at 55˚C, 3 min at 68˚C, then a final 7 min at 68˚C. The cDNA library was purified using AMPure XP Beads (Beckman Coulter), according to manufacturer's instructions. Finally, the cDNA library was visualised using Agilent 2100 Bioanalyser, High Sensitivity DNA chip, and quantified using Qubit® 2.0 Fluorometer, under standard operating procedures.

2.2.8.3 Illumina TruSeq stranded total RNA library preparation

Whole RNA libraries were prepared using TruSeq v2 RNA-seq Library Preparation kit (Illumina), according to the manufacturer's instructions. Following rRNA removal (Epidemiology kit, Epicentre) RNA was fragmented using Elute, Prime, Fragment High Mix and a 1:1000 dilution of ERCC Synthetic Spiking Standards (LifeTechnologies) was added prior to an 94˚C incubation for 8min. First and second strand cDNA synthesis was performed by two consecutive steps; First stand synthesis Act D and Superscipt II Reverse Transcriptase were added prior to thermocyler incubation of 10 min at 25˚C, 15 min at 42˚C and 15 min at 70˚C followed by addition of the second strand marking mix and 1-hour incubation at 16˚C. The cDNA was purified using AMPure XP beads (Beckman Coulter), according to manufacturer's instructions, prior to 3' end adenylation using A-tailing mix. The index adaptors were ligated to the cDNA strands by incubation at 30˚C for 10 min in the presence of ligation mix and RNA adapter indexes. Excess adaptors were removed using AMPure XP beads (Beckman Coulter), according to manufacturer's instructions. The libraries were subsequently amplified by PCR under following conditions: 30 sec at 98˚C, 15 cycles of 10 sec at 98˚C, 30 sec at 60˚C, 30 sec at 72˚C, then a final 5 min at 72˚C. The amplified libraries were purified using AMPure XP Beads (Beckman Coulter), according to manufacturer's instructions, prior to assessing quality using Agilent 2100 Bioanalyser, DNA 1000 chip, and quantified using Qubit® 2.0 Fluorometer, under standard operating procedures.

2.2.9 RNA sequencing data analysis

2.2.9.1 RNA sequencing Alignment

The RNA sequencing results were analysed and validated in collaboration with colleagues from the GSTT&KCL BRC Bioinformatics core. All steps were performed using in-house Python scripts and the R packages FastQC and Tuxedo suite. Quality of RNA sequencing data was assessed using FastQC prior to alignment to known transcripts and the reference genome - including ERCC

synthetic controls - using TopHat2. The aligned RNA fragments were then assembled into transcripts using Cufflinks and CuffMerge generating count and FPKM (Fragments Per Kilobase of Exon Per Million Fragments Mapped) values per transcript. Counts were normalised against library size and filtered for a minimum read-count of n=smallest sample group e.g. n=24 controls whereas FPKM values were normalised against library and transcript size.

2.2.9.2 Principle components analysis (PCA)

Phenotype data file and normalised count file were loaded into R. Counts per million (CPM) were calculated and principle component analysis (PCA) performed. The table with calculated PCA values was saved and plotted using various pairs of principle component values (PC1 vs PC2, PC2 vs PC3, PC4 vs PC5) and coloured-coded for disease type, batch, sex or age using ggplot (R) (**see Appendix 1 for the script**).

2.2.9.3 Differential expression analysis

Differential expression between phenotype groups was assessed using EdgeR (Bioconductor). Prior to this the 'remove unwanted variables' (RUV) method RUVseq (Bioconductor) was employed and principle components analysis (PCA) using ggplot2 (R) to assess any unwanted variation and biases in the data. EdgeR used TMM (trimmed mean of M) with M=library size, to correct the count values for RNA library compositional biases prior to building a matrix incorporating normalised count values, RUV values and PCA values for age, sex, batch and disease type. EdgeR then assessed any difference in expression due to disease status following correction of all other observed biases within the data. The rate of type I error was controlled for using the Benjamini-Hochberg method for false discovery rate (FDR). Output files contained fold change, p-value, FDR controlled q-value, and counts per million (CPM) (**See Appendix 2 for the script)**.

To confirm pathway analysis results, the differential expression analysis was repeated for CD vs controls now including smoking status as a covariate in addition to above mentioned covariates.

2.2.9.4 Gene Set Enrichment Analysis

Gene Set Enrichment analysis (GSEA) (https://www.broadinstitute.org/gsea - Broad Institute) [148] was employed to assess enrichment of known biological pathways within the differentially expressed genes. GSEA assesses the presences of the genes of interest against known gene sets within their Molecular Signatures Database (MSigDB) [148]. A pre-ranked analysis method was employed within GSEA where a ranking score was calculated for each gene by multiplying the –LOG10(q-value) by the direction of fold change (e.g. -1 or 1), listing genes with a highest significant change in expression and a positive or negative fold change at the top or bottom of the ranked list, respectively. Further parameters allowed up to 1000 permutations and applied a weighted enrichment statistic. GSEA examined if members of MSigDB gene sets were clustered towards the top or bottom of the ranked DE gene list, determining correlation between the gene set and the phenotypic class by weighting each step. Statistical methods calculated a Normalised Enrichment Score (NES) for each gene set incorporating the significance of the Enrichment Score (ES). In addition, type I error was controlled for by calculating a false discovery rate (FDR).

A two-sided probability test was employed using the “prop.test” function in R, to assess if the probability of success in several groups was the same by calculating the chi-square.

2.2.9.5 Ingenuity Pathway Analysis

Ingenuity pathway analysis (IPA) core analysis (IPA®, QIAGEN Redwood City, www.qiagen.com/ingenuity) was utilised to investigate if genes involved in curated canonical biological pathways were affected by the previously identified differentially expressed gene lists. IPA employs the ingenuity knowledge base, a platform built upon a wide range of biological information including textbooks, reviews, published biomedical literature, internally curated knowledge and a variety of public databases. A spreadsheet containing ensemble IDs, fold change values, q-values and expression (FPKM) values was

uploaded into IPA. All genes identified to be expressed with intestinal tissue (FPKM >1, n=15,553) were indicated as the reference set and genes that were significantly differentially expressed (DE q-value <0.05) were indicated as genes of interest. Furthermore, only effects observed within IPA pathways curated from data collected from humans and human cell lines were included. IPA calculates a p-value using the right-tailed Fisher Exact Test and a ratio score for each pathway where, the p-value indicates the likelihood that the association between input genes and a pathway is due to random chance and ratio being based on the number of input genes per pathway over the total number of genes in that pathway.

2.2.9.6 Expression quantitative trait loci (eQTL) analysis

Matrix eQTL (R) (http://www.bios.unc.edu/research/genomic_software/Matrix_eQTL/runit.html) was employed to perform a genome wide eQTL analysis, to explore correlations in variation of gene expression within the colonic transcriptome with genetic variation in each individual (SNPs). The input data included quantitative gene expression values (FPKM), genome-wide SNP genotype data for 241,995 SNPs from Infinium Human Core Exome or Human Core array and optional additional covariate data, (see section 2.2.9.6.1 – 2.2.9.6.3 for details). The analysis parameters within Matrix eQTL were set to look for significant associations (or correlations) between each of 241,995 input SNPs with changes in expression of genes located within 1Mb of the SNP (*cis*-eQTLs). P-values and False Discovery Rate (FDR) values were generated by Matrix eQTL using a linear regression and the Benjamini-Hochberg method respectively (**See Appendix 3 for the script)**.

2.2.9.6.1 Gene expression data input

Generated colonic whole RNAseq transcriptome data was utilised, including non-coding RNAs, to perform a comprehensive eQTL survey. A gene expression matrix was created including all transcripts with expression above 1 FPKM (Fragments Per Kilobase of exon per Million fragments mapped), for each of

the 127 patient or control samples. Transcripts located on chromosomes X and Y were excluded, resulting in a total of 17,258 genes. Prior to entry into Matrix eQTL FPKM values were standardized to have a mean of zero by dividing the FPKM of a gene minus the average FPKM of that gene by the standard deviation.

2.2.9.6.2 Genotype data input

Generated genome-wide SNP data from 121 samples was utilised to perform the eQTL survey. Genome studio was used to export raw data from the samples prior to processing by PLINK to remove indels, tri-allelic SNPs, SNPs with a minor allele frequency below 5% and SNPs not within Hardy-Weinberg Equilibrium ($p<0.0001$). Due to the 2 different SNP arrays used there was incomplete data for a subset of SNPs, only SNPs containing genotype data for over half of the samples ($n \geq 60$) were included in the analysis. The genotype file following filtering of all above mentioned criteria contained genotype data for 241,995 SNPs.

2.2.9.6.3 Covariates input files

To enable correction for any possible bias in the expression data, a covariate file containing the principle components 1-4 from the original PCA analysis (see section 2.2.9.2) was included in the matrix eQTL analysis.

2.2.9.6.4 IBD associated SNP coverage

Linkage disequilibrium (LD) was assessed between 502 IBD index SPNs (based on current literature) and the genome-wide SNP genotype data for 241,995 SNPs evaluated within the patient and control samples using Infinium Human Core Exome or Human Core array (see section 2.2.7), in order to establish coverage of IBD associated SNPs within the eQTL analysis. The 1000 Genomes phase 3 reference data were utilized to calculated LD scores using PLINK. Files containing the genotyped SNPs and IBD index SNPs including their genome locations were loaded into PLINK and the reference data set was selected as the

1000 Genomes phase 3 reference data, prior to identifying direct matches between the IBD index SNPs and genotyped SNPs. Following removal of the direct matched SNPs, LD scores were calculated for each IBD index SNP with the genotyped SNP. Any genotyped SNP with an $r^2 \geq 0.8$ was indicated as tagging an IBD index SNP.

2.2.10 Intestinal tissue cell phenotyping and analysis

2.2.10.1 Generation of single cell suspension from intestinal pinch biopsies

Single cell suspensions were generated from 2-4 uninflamed gut pinch biopsies per patient collected during routine colonoscopies at Guy's and St Thomas' Hospital after informed consent was taken. The biopsies were transferred into 10 ml collection media and incubated for 10 min at 37˚C under shaking conditions. The tissue was transferred into a petri-dish and cut into tiny pieces using forceps and a scalpel. Next, the pieces were washed off the petri-dish into a falcon tube containing 10 ml of freshly prepared digest followed by an incubation at 37˚C for 30 min under shaking conditions. The remaining connective tissue bonds were disrupted by syringe aspiration through a 30-gauge needle. The single cell suspension was collected in 50 ml falcon tube and debris was removed by passing cells through a 100 μm cell strainer (BD). Subsequently, cells were washed and resuspended in culture media prior to counting and estimating viability using Trypan blue dye exclusion (SIGMA).

2.2.10.2 Flow cytometry based cell phenotyping

The generated single cell suspensions (n=25) were distributed at $2x10^5$ cells/tube in 5 ml polystyrene tubes (n=5 tubes per donors; unstained, antibody mix 1, 2, 3 and 4) prior to incubation on ice for 10 min in the presence of FcR block (Miltenyi) followed by 30 min in the presence of the relevant mAB cocktail (**Table 2.3 -Table 2.5**). After washing with FACS buffer and 1 μM DAPI (AnaSpec) was added prior to analyses by FACS CantoII.

Table 2.4 | Biopsy antibody cocktail

Antibody	Amount (μl)	Supplier	Cell type
CD326 – PE	2.5	BioLegend	Endothelial cells
CD45 – APC/Cy7	2.5	BioLegend	Lymphocytes
DAPI	-	AnaSpec	Live/Dead stain

Table 2.5 | FMO for CD45 with Isotype control

Antibody	Amount (μl)	Supplier	Cell type
CD326 – PE	2.5	BioLegend	Endothelial cells
APC/Cy7-IgG k1 Mouse	2.5	BioLegend	-
DAPI	-	AnaSpec	Live/Dead stain

Table 2.6 | Leukocyte antibody cocktail

Antibody	Amount (μl)	Supplier	Cell type
CD45 – APC/Cy7	2.5	BioLegend	Lymphocytes
CD4 – APC	2.5	BioLegend	T helper cells
CD8 – FITC	2.5	BioLegend	Cytotoxic T cells
CD14 – PE	2.5	BiolLegend	Monocytes
CD68 – PE-Cy7	2.5	BioLegend	Macrophages
CD66b – PerCP/Cy5.5	2.5	BioLegend	Neutrophils
DAPI	-	AnaSpec	Live/Dead stain

2.2.10.3 Deconvolution biopsy composition

A deconvolution model was designed and build to enable prediction of cell type fractions within the intestinal biopsies from which gene expression data was generated (see section 2.2.2). Cellular phenotype data generated by flow cytometry (see section 2.2.10.2) for n = 21 was loaded into SAS (statistical analysis software) and combined with normalised count expression data for 15,517 genes, for these patients. A univariate analysis using a marginal model was employed to identify which genes exhibited the most significant influence on cell count. The Mixed Model included intra-patient covariance to account for the known correlation between the cell types (using SAS). The genes identified to be significantly associated ($p \leq 3.2x10^{-6}$) with cell count through the univariate analysis were progressed into a multivariate analysis to identify

which set of significant genes can collectively predict cell count per cell type, using lasso (least absolute shrinkage and selection operator). The univariate analysis in lasso was performed for each cell type individually, using the same set of significant input genes. Cell types predicted by gene expression included epithelium cells ($CD45^{neg}/CD326^{pos}$), leukocytes ($CD45^{pos}$), T helper cells ($CD45^{pos}/CD4^{pos}$), cytotoxic T cells ($CD45^{pos}/CD8^{pos}$) and monocytes ($CD45^{pos}/CD14^{pos}$). Lasso generated an estimate score for each of between 16-20 'predictive genes', in addition to an intercept score for each cell type (**Appendix 6**). The cell count for each cell type was calculated by multiplying the gene scores by the gene expression value for each of the 'predictive genes' and added to the intercept value (see equation below).

$$Cell\ type = \sum (Gene\ estimate\ score * gene\ exprression\ value\ ) + intercept\ value$$

Using the estimate model that was constructed, predicted count values were compared with observed values (generated by flow cytometry) for the n = 21 patient used to build the model (using SAS). Finally, the estimate model was employed to predict cell counts for 57 CD intestinal biopsies based on their gene expression values of the 'predictive genes'.

2.3 Methods Project 2: Investigation of transcriptional biomarkers in prediction of relapse

2.3.1 Power Calculation

This study was designed as a pilot study. Under the current design, identifying transcriptional signatures between relapsed and non-relapsed CD patients within three different patient groups, we are not powered to identify differentially expressed genes with small effects or which are high variability. It was decided to recruit CD patients from three different cohorts; i) Post-

resection surgery patients, ii) patients being withdrawn from anti-TNF treatment, or iii) routine gastroenterology/IBD clinic patients who have had active disease within the last year, for logistical reasons. If the generated data form this pilot study supports our hypothesis that changes in the transcriptome can indicate/predict relapse in CD patients, a full size study should be initiated.

2.3.2 Sample Collection

Blood samples were collected, either by myself or Dr Kamal Patel, from CD patients at Guy's and St Thomas' Hospital after informed consent was given. Recruited patients included 49 patients in endoscopic remission (Rutgeerts score $\leq$ 1) or self-reported clinical remission at time of recruitment. Their progress was monitored over a 12-month period to determine who did and did not relapse. A peripheral blood samples (50 ml) was collected in EDTA tubes (BD Biosciences) for isolation of peripheral mononuclear cells (PBMCs) at time of recruitment to the study (remission sample) and either at time of relapse or after 12 months, whichever came first (follow-up sample). Patients within three subsets were recruited:

- 20 post-surgery patients, either resection or stoma reversal surgery.
- 20 routine gastroenterology clinic patients
- 9 patients having been withdrawn from anti-TNF treatment, Humira or Infliximab

The post-surgery patients were recruited either two weeks post their resection surgery or at the day of their stoma reversal, guaranteeing full remission. The routine gastroenterology clinic patients were recruited when reporting clinical remission during their appointment and the anti-TNF withdrawal patients were recruited eight week following their final Infliximab injection (or two weeks for Humira). Relapse was assessed at a 6-month follow-up colonoscopy for the post-surgery patient cohort, with a Rutgeerts score $\geq$ 2 classified as relapse. For the anti-TNF withdrawal and gastroenterology clinic patient cohorts, relapse was classified as a need to change the patients' medication. Mean age for the

Gastro-clinic patients was slightly higher, at 44.8, than for the post-surgery (32.7) and anti-TNF withdrawal patients (33.3) (**Table 2. 7**).

Table 2. 7 | Patient demographics

	Post-surgery patients	Anti-TNF withdrawal patients	Gastro-clinic patients
Mean age	32.7	33.3	44.8
# Female	14	5	8
Percentage	70%	56%	40%
# Male	6	4	12
Percentage	30%	44%	60%

2.3.3 PBMC isolation

Whole blood was transferred to 50 ml falcon tubes and under-layered with 10 ml lymphocyte separation media (MP Biochemicals Europe) prior to centrifugation (2000 rpm, 20 min, 21˚C) without brakes. Following centrifugation, lymphocytes were isolated from the interphase, transferred to a new 50 ml Falcon tube and washed with PBS (2000 rpm, 10 min, 21°C). Subsequently, cells were washed twice more with PBS (1000 rpm, 10 min, 21°C) and resuspended in PBMC media prior to counting and estimation of viability using Trypan blue dye exclusion (SIGMA). Isolated PBMCs were cryopreserved in FBS + 10% DMSO, using temperature controlled freezing at -80˚C.

2.3.4 Thawing and resting of cells

Cryopreserved PBMCs were thawed in a 37˚C water bath prior to dilution in 12 ml of cold (4°C) PBMC media. Following centrifugation at 1800 rpm for 5 min, cells were washed twice with 15 ml PBMC media (5 min at 1800 rpm). Supernatant was removed and cells were resuspended at $1x10^6$ cells/ml in PBMC media + 10U/ml DNAse I and transferred into a cell culture flask (ThermoFisher) prior to 2-hour incubation at 37˚C and 5% CO_2.

2.3.5 Flow cytometry based cell sorting

Thawed and rested PBMCs were transferred into 50 ml Falcon tube following detachment of cells from cell culture flask using cell scraper and cold PBS washes. Cells were pelleted (5min at 1800 rpm, 4˚C), resuspended and counted using Trypan Blue Dye exclusion. Cells were washed in FACS buffer and incubated for 10 min in the dark on ice in the presence of FcR block followed by 30 min in the presence of the cell sorting mAB cocktail (**Table 2. 8**). After washing with FACS buffer, the cells were filtered through a 30 μm grid (Miltenyi) and resuspended at 10-20 x10^6 cells/ml in cell sorting solution. 1 μM DAPI (AnaSpec) was added prior to cell sorting into $CD4^{pos}$, $CD8^{pos}$ and $CD14^{pos}$ cell population using an Aria III cell sorter (BD) which was performed by the Guy's and St Thomas/KCL BRC Flow Cytometry Core Service. The purity of the separated cell populations, $CD4^{pos}$, $CD8^{pos}$ and $CD14^{pos}$, was then assessed by Flow Cytometry.

Table 2. 8 | Cell sorting antibody cocktail

Antibody	Amount (μl)	Supplier	Cell type
CD3 – PerCep/Cy5.5	5.0	Cambridge Biosciences	T cells
CD14 – PE	5.0	Cambridge Biosciences	Monocytes
CD4 – APC	5.0	Cambridge Biosciences	T helper cells
CD8 – FITC	5.0	Cambridge Biosciences	Cytotoxic T cells
DAPI	-	AnaSpec	Live/Dead stain

2.3.6 Cell culture and activation

Following cell separation (section 2.3.4), purified populations of immune cells ($CD4^{pos}$, $CD8^{pos}$ or $CD14^{pos}$) were pelleted via centrifugation (5 min at 1800 rpm, 4˚C), supernatant removed and resuspended in PBMC media at 2x10^5 cells per well (at 2 wells per sample) (**Figure 2.1**). 1 out of 2 wells per sample were immune stimulated by either CD3/CD28 T-activator beads (ThermoFisher) at 50 μl per million cells, ($CD4^{pos}$ and $CD8^{pos}$ cells) or 0.5 mg Lypopolysaccaride (LPS) (Sigma-Aldrich) ($CD14^{pos}$ cells) prior to 4-hour

incubation at 37˚C and 5% CO_2. For samples with less than $2x10^5$ cells per sample no stimulation was executed. Following incubation, cells were transferred into Eppendorf tubes and centrifuged at 300 x g for 5 min at 4˚C. Supernatant was removed and cell pellets resuspended in 700 μl Qiazol (Qiagen) followed by a 1 min vortex to lyse cells. Lysates were stored at -80˚C until further processing.

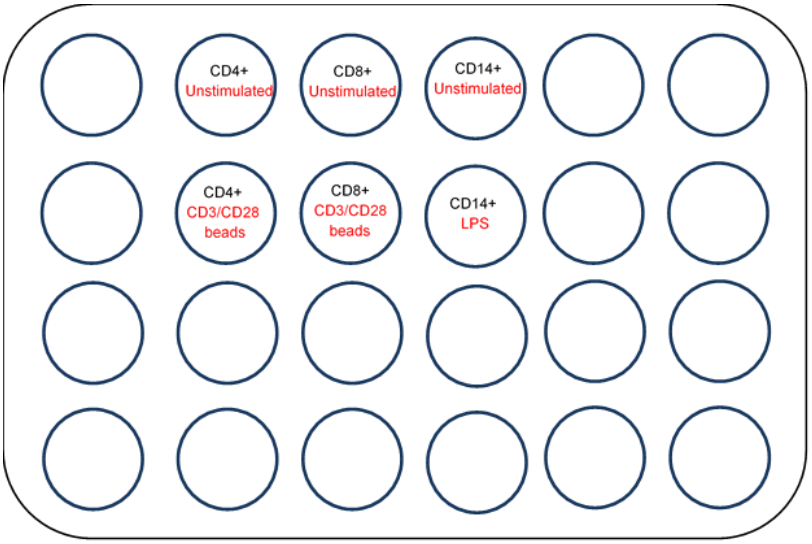


Figure 2.1 | Cell culture plate layout
Example cell culture plate, with each circle representing a well, containing $CD4^+$, $CD8^+$ and $CD14^+$ separated cells, stimulated and unstimulated.

2.3.7 RNA preparation

RNA was extracted using an RNeasy miRNA micro Kit according to manufacturer's instructions (QIAGEN Gmbh). Briefly, the samples were thawed at 37˚C, incubated for 5 min at room temperature prior to addition of 180 μl chloroform. Following 15 min centrifugation at 4˚C and 12,000 x g, the aqueous phase was transferred and ethanol added prior to transfer onto MiniElute columns. Multiple wash steps were performed on column using RWT and RPE buffer. Genomic DNA was removed using RNase free DNaseI (Qiagen). Lastly, 28 µl RNase-free water was added during two elution steps, to elute the RNA. RNA was quantified using Qubit fluorimeter.

2.3.8 Real-time qRT-PCR

2.3.8.1 Reverse transcription polymerase chain reaction (RT-PCR).

30 ηg of RNA was added to 4 μl 5x iScript reaction mix (containing oligo T primers), 1 μl iScript reverse transcriptase enzyme (iScript cDNA Synthesis Kit, BIORAD) and nuclease-free water. After gently mixing samples were incubated at 25°C for 5 min followed by 30 min at 42 °C and finally 5 minutes at 85 °C. The resultant cDNAs were stored at -20°C until needed for further steps.

2.3.8.2 Assessing cell stimulation by Real-Time quantitative PCR (RT-qPCR) of the TNFα gene

25 ng cDNA was dispensed in a 96-well plate along with qPCR master mix (ABgene), as well as Taqman assay primers for either the TNFα target gene or 18S endogenous control gene (**Table 2.2**). Final reaction volume was made up to 20 μl with RNAse free water. Following centrifugation, the plate was placed in the ABI 7900-HT Real-Time PCR system and amplified under the following conditions: 15 min at 95°C, 40 cycles of (15 sec at 95°C, 1 min at 60°C). Abundance of mRNA was calculated by comparing the threshold PCR cycle (Ct, cycle at which logarithmic amplification was observed) of the unstimulated samples with the stimulated samples, for both the target and the endogenous control gene (RP18S) and represented as ΔCt (delta cycle threshold). The relative quantification (RQ) was then calculated by taking 2^(ΔCt PR18S - ΔCt TNFα).

2.3.9 MicroArray sample preparation

Amplification and labelling of low input RNA samples for gene expression microarray analysis was performed by Dr David Chambers from the Drug Discovery Unit, Wolfson CARD, King's College London. See the methods described below.

2.3.9.1 NuGen Ovation RNA Amplification System V2

RNA extracted from $CD4^{pos}$, $CD8^{pos}$ and $CD14^{pos}$ separated cells - stimulated and unstimulated - were normalised to 1 ng/μl and quality assessed using

Bioanalyser Pico RNA chips (Agilent technologies). 5 ng RNA starting material per sample was used for cDNA generation and amplification using Ovation RNA amplification system V2 (NuGen), according to manufacturer's protocol.

2.3.9.2 NuGen Encore BiotinIL module

Above mentioned amplified cDNA was labelled for hybridization to HumanHT-12 V4 expression BeadChip (Illumina) using Encore BiotinIL module (NuGen), according to manufacturer's protocol, generating 2-4 μg of labelled cDNA per sample. Samples were quantified using Nanodrop One/One^{C} (Thermo Scientific) and normalised to 150 ng/μl prior to submission to the BRC Genomics core for microarray analysis using the HumanHT-12 V4 expression Beadchips (Illumina).

3. Quality control colonic transcriptomics data

3.1 RNA and DNA quantity and quality

An average of 60μg of total RNA was extracted from 20-40mg of large intestinal tissue with RNA Integrity Number (RIN) values ranging from 2.2 to 8.6, the average RIN being 7. The Agilent software tool calculated RIN scores per sample based on the ratio of the ribosomal bands as well as the presence or absence of degradation product. RIN values range from 1-10, with scores of >7 indicative of high quality RNA. For this work, a cut-off of 6 was implemented as the RNA sequencing technology does not rely on poly(A) binding. Overall 159 bowel tissue samples were collected and 134 yielded RNA with a RIN >6. The absence of DNA contamination was confirmed using a reverse transcriptase (RT) negative control during the amplification of the cDNA sequencing libraries. Additionally, 175 blood samples for RNA (Paxgene tubes) and DNA extraction (EDTA tubes) were collected in parallel to the biopsies. Paxgene whole blood RNA extractions were performed on 151 samples resulting in a mean yield of 5.9 μg of whole RNA and a mean RIN of 7.5. DNA extraction resulted in a mean yield of 153 μg, and was completed for 174 out of 175 samples. For this study 128 high quality RNA and matched DNA samples were taken forward for whole RNA sequencing and genotyping.

3.2 Ribosomal depletion and library preparation

RNA samples to be used for Next Generation Sequencing (NGS) of the whole transcriptome, require depletion of ribosomal RNA (rRNA). Ribosomal depletion relies on rRNA specific probe binding prior to removal using magnetic beads, preserving all coding and non-coding transcripts. Intestinal biopsy samples contained a high bacterial content, therefore a combined RiboZero kit depleting both human and bacterial rRNA (see Materials and Methods 2.2.8.1) was used. Bioanalyzer electropherograms were generated pre- and post- rRNA depletion to confirm effective removal of 18S and 28S rRNA peaks (**Figure 3.1**) which was a requirement for taking the samples forward for library preparation.

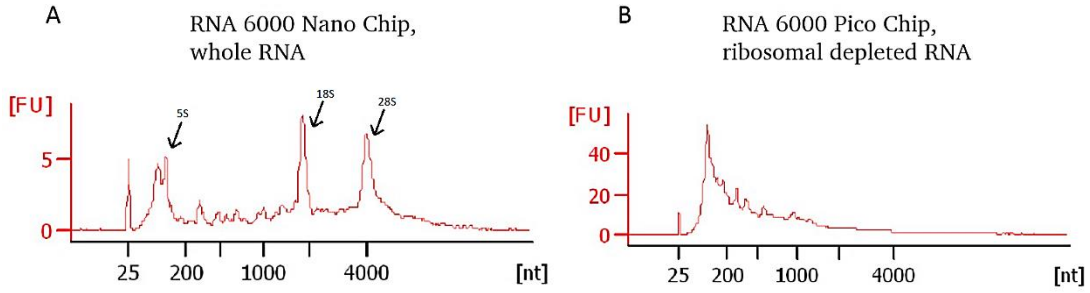


Figure 3.1 | Quality control RNA ribosomal removal
Electropherograms showing fluorescence units (FU) on y axis versus size, nucleotides (nt), on the x axis. **(A)** Total RNA; marker (25nt), miRNA and 5s rRNA (<200nt), 18s subunit (~2000nt) and 28s subunit (~5000nt). **(B)** Ribosomal depleted RNA; marker (25nt) and mRNA + miRNA (100 – 2000nt). (Plots generated through the bioanalyzer (Agilent Technologies)).

The two main subunits of human ribosomal RNA are 18S and 28S with sizes 1869 nt and 5070 nt, respectively. The 16S and 23S bacterial ribosomal subunits could not clearly be identified, suggesting low abundance of bacterial rRNA or a lack in sensitivity of the RNA Nano chip. The bacterial and human 5S subunit (120 nt) would most likely be present in the peak observed at <200 nt, co-localising with the miRNA fraction (**Figure 3.1A**). Post-ribosomal depletion bioanalyser traces suggested complete elimination of the previously observed ribosomal peaks, resulting in whole RNA between 150-2000 nucleotides (nt) (**Figure 3.1B**). cDNA libraries for sequencing were subsequently prepared; 32 libraries were generated using the Epicentre Scriptseq chemistry and 96 using the Illumina Truseq chemistry. The change in chemistry was decided upon following persistent issues with the Epicentre ScriptSeq kit and poor reproducibility. Following the library preparation and clean up the quality of the cDNA libraries was assessed via analysis on a DNA high sensitivity or DNA 1000 bioanalyzer chip for the Epicentre ScriptSeq and Illumina TruSeq cDNA generated libraries, respectively (**Figure 3.2**).

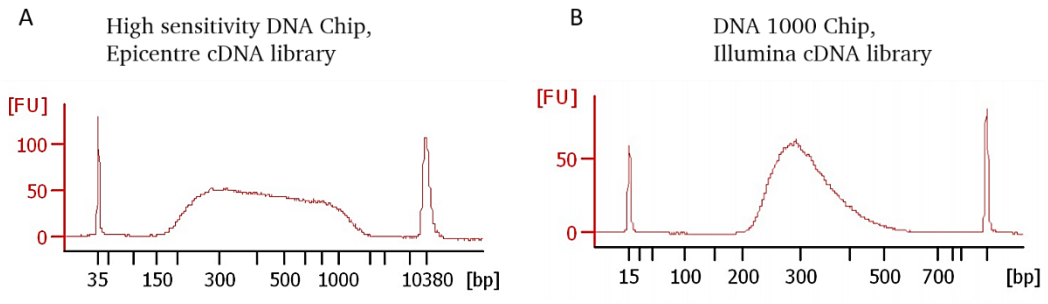


Figure 3.2 | Quality control library preparation
Electropherograms showing fluorescence units (FU) on y axis versus size, base pairs (bp), on the x axis **(A)** cDNA library (150–1000bp) and markers following library preparation using the Eppicentre Scriptseq kit. **(B)** cDNA library (200–500bp) and markers following library preparation using the Illumina Truseq kit. (Plots generated through the bioanalyzer (Agilent Technologies)).

The median of cDNA library size was observed, as expected, at approximately 200-300 bp (100 bp fragments plus ~60-70 bp adapters ligated on either side). The Illumina TruSeq libraries showed high symmetry within their distribution and high similarity across cDNA library samples (**Figure 3.2B**) whereas the Epicentre ScriptSeq libraries showed wider variation in distribution, between 200 and 1000, with an extended plateau like shape (**Figure 3.2A**) and larger inter-sample variation. The observed variation within the Epicentre Scriptseq cDNA libraries suggested lower efficiency during RNA fragmentation.

3.3 FastQC – RNA sequencing QC

Quality control of the RNA sequencing data was performed by the BRC bioinformatics unit using FastQC. FastQC calculated a quality score (Q) of each given base by estimating the probability of the base being called incorrectly. A Q of 20 represents an error rate of 1 in 100 and a Q of 30 an error rate of 1 in 1000 corresponding with a base call accuracy of 99% and 99.9%, respectively. The FastQC plots (**Figure 3.3A-D**) represent the highest and lowest quality 100 bp cDNA library fragments using either Epicentre or Illumina library preparation methods. Libraries created with the Epicentre chemistry showed good quality overall, with the majority of base pairs showing Q ≥30. However, variability in quality, specifically towards the end of the library fragment, was observed within the Epicentre Libraries (**Figure 3.3 A.1-A.2**). FastQC plots of

cDNA library fragments generated using the Illumina chemistry are indicative of consistent high quality sequencing, with Q scores approaching 40 (**Figure 3.3 B.1-B.2**). Overall, Figure 3.3 suggests the cDNA libraries have been sequenced to high accuracy.

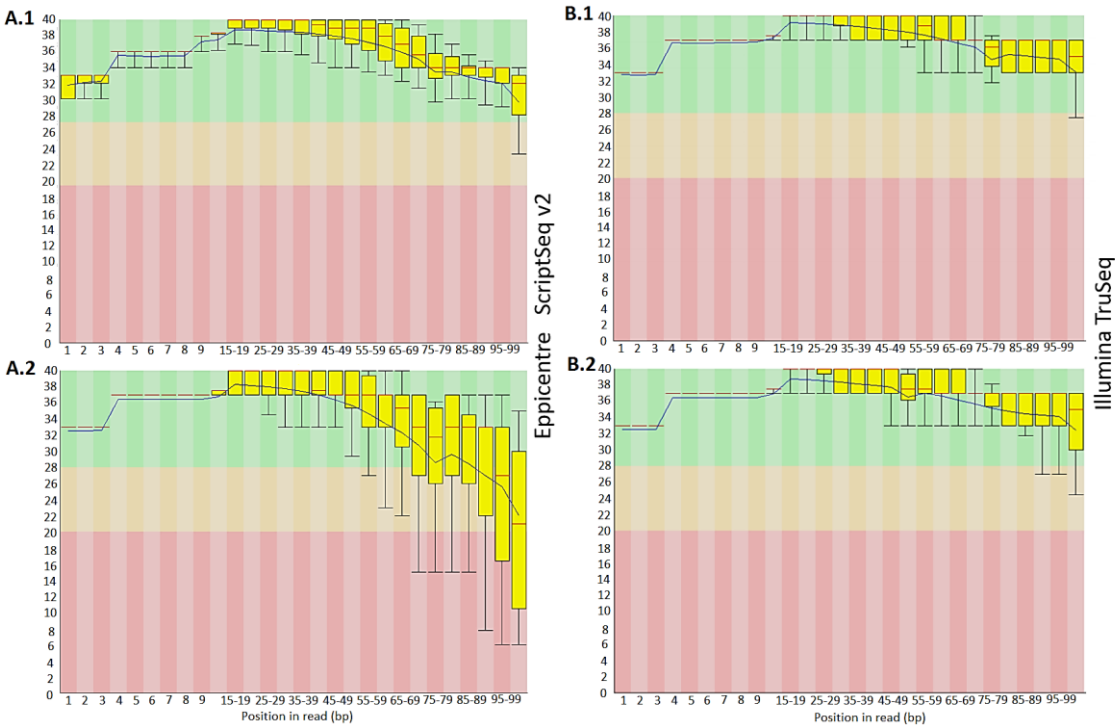


Figure 3.3 | FastQC plot
Quality score (Q), calculated by estimating the probability an incorrectly called base shown for each position (bp) in a cDNA fragment read. Within Epicentre generated libraries (A) and Illumina generated libraries (B). Where A.1 and A.2 represent the best quality libraries and A.2 and B.2 the lowest quality libraries.

3.4 RNAseq read alignment

Following the quality check by FastQC, the Tuxedo Suite was employed for further analysis of the RNA sequencing data. The alignment was performed by the BRC bioinformatics unit. Tophat2, part of the Tuxedo suite, enabled alignment of the RNA-sequence reads against known transcripts and the reference genome, providing the most likely genomic location from where the sequence reads originated. Following alignment, TopHat2 estimates the

percentage of reads per sample mapped back to the reference transcriptome and reference genome, allowing evaluation of the data (**Figure 3.4**). The 32 Epicentre ScriptSeq cDNA libraries were sequenced in two batches of 16 samples, P632 (**Figure 3.4A**) and P344 (**Figure 3.4B**), with 4 indexed and pooled samples run across 2 lanes of a flow cell, in order to provide a target coverage of 100X. The 96 Illumina TruSeq cDNA libraries were sequenced in three batches of 32 samples, P478 (**Figure 3.4C**), P524 (**Figure 3.4D**) and P566 (**Figure 3.4E**), at 4 samples per lane.

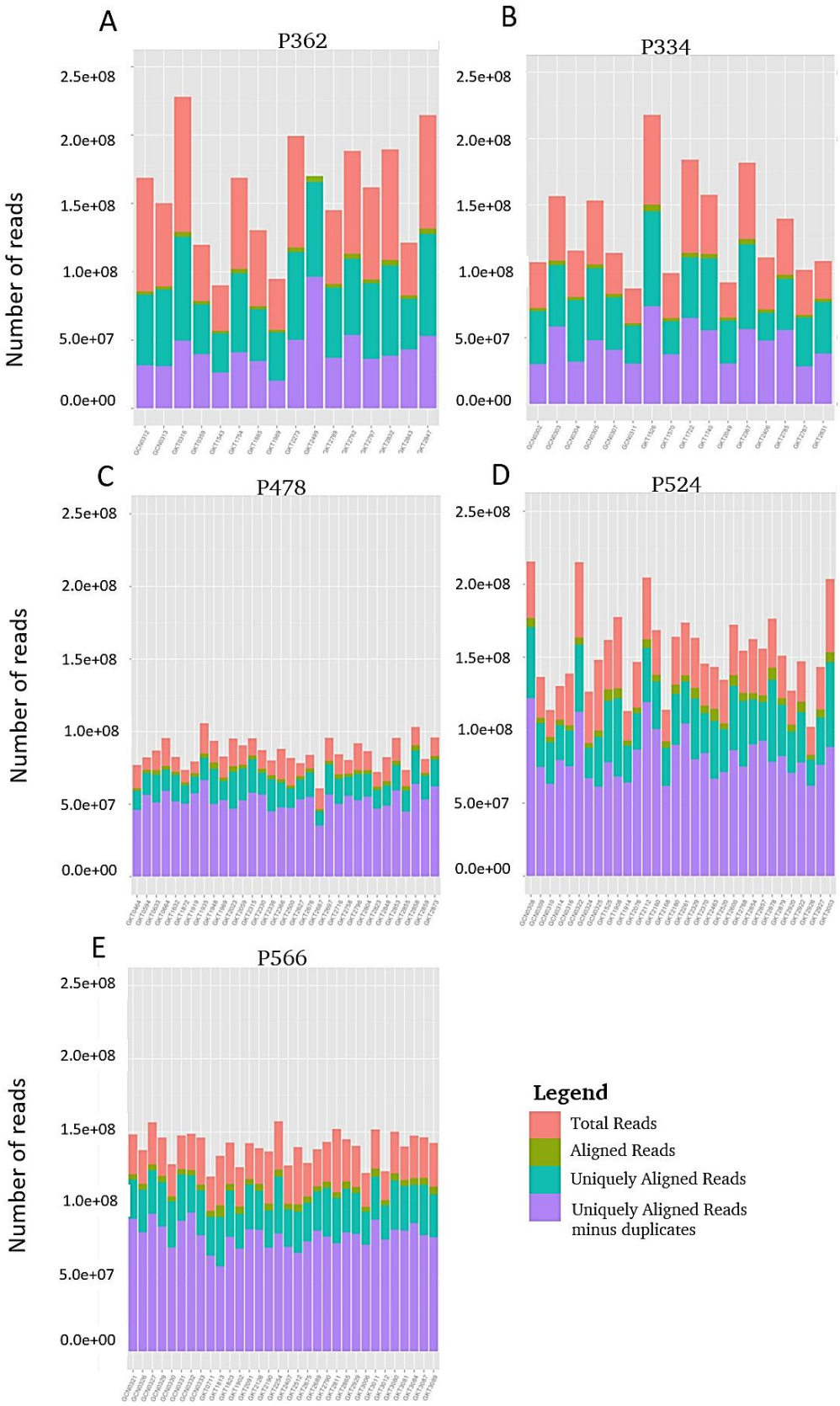


Figure 3.4 | RNA sequencing read alignment per sample
Stacked column chart displaying number of RNA sequencing reads aligned to the reference genome (y-axis) per individual sample (x-axis) in each sequencing batch (A-E). Height of bar shows total number of reads sequenced (red), total number of aligned reads (green), number of uniquely aligned reads (turquoise) and number of uniquely aligned reads minus duplicate reads (purple). (Alignment was performed by the Bioinformatics unit)

Generated transcription reads showed a high level of alignment to the reference (60-70%: P362 and P334; 80-90%: P478, P524 and P566) (**Figure 3.4**). Considering the libraries contained transcripts from both non-coding and coding RNAs, making mapping more challenging, the observed mapping rates are indicative of good performance of the RNA sequencing experiment. Following removal of duplicate reads (30%-50%), a coverage of approximately 40 million reads for each of the Epicentre libraries and 70 million reads for the Illumina libraries was achieved (**Figure 3.4**). Transcripts were assembled using Cufflinks and CuffMerge, generating datasets containing count values or FPKM (Fragments Per Kilobase of exon per Million fragments mapped) values for further analysis.

3.5 ERCC-Spike in controls

In addition to the QC steps performed during analysis, an ERCC spike-in control was added into the libraries prior to sequencing to provide an internal control. Known concentrations of ERCC spike-in control RNAs were plotted against output FPKM values to produce a standard curve which showed high correlation between input amount and FPKM estimate ($R^2 = 0.907$) (**Figure 3.5**). Once high correlation was established, the ERRC spike-in controls were utilized to correct for any biases between libraries. Covariate values were calculated based on ERCC spike-in control values and incorporated in any downstream analysis (Materials and Methods).

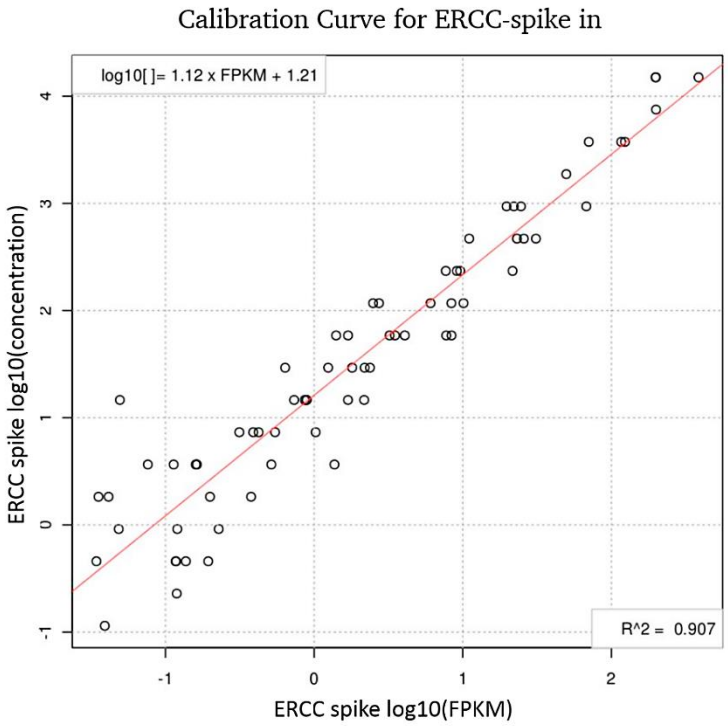


Figure 3.5 | Calibration curve of ERCC spike-in control
Calibration curve showing the ERCC spike concentrations versus measured FPKM spike values, with $R^2 = 0.907$.

3.5 Principle component analysis

As a final check, following successful alignment, removal of duplicates, transcripts assembly and QC, the presence of potential biases in the data was investigated. Principle component analysis (PCA) was performed to assess variation between samples based on the generated normalised gene expression data. PCA is a statistical procedure which converts a set of potentially correlated variables into a set of linearly uncorrelated variables called principle components (PC). Multiple principle components can be generated in this way with the first principle component (PC1) having the largest possible variance. The first six principle components were investigated to establish if the observed variance in the dataset was influenced by any of the following potential covariates: disease status, sex, age and sequencing batch (**Figure 3.6** and **Figure 3.7**).

When plotted, PC1 versus PC2 indicated no effect due to sex (**Figure 3.6B**) or age (**Figure 3.6C**). For disease status, a minor effect of phenotype on distribution was observed with CD cases clustering slightly further to the lower left of the plot than UC or controls (**Figure 3.6A**). When investigating the PCA distribution with respect to sequencing batch a clear clustering together of batches P334 (red) and P362 (olive green) versus P478 (green), P524 (blue) and P566 (purple) was observed (**Figure 3.6D**). This demonstrated a clear bias related to the sequencing chemistry used; P334 and P362 were generated using the Epicentre ScriptSeq kit whereas P478, P524 and P566 were generated using the Illumina TruSeq kit. Correction for this effect was therefore implemented within downstream analysis.

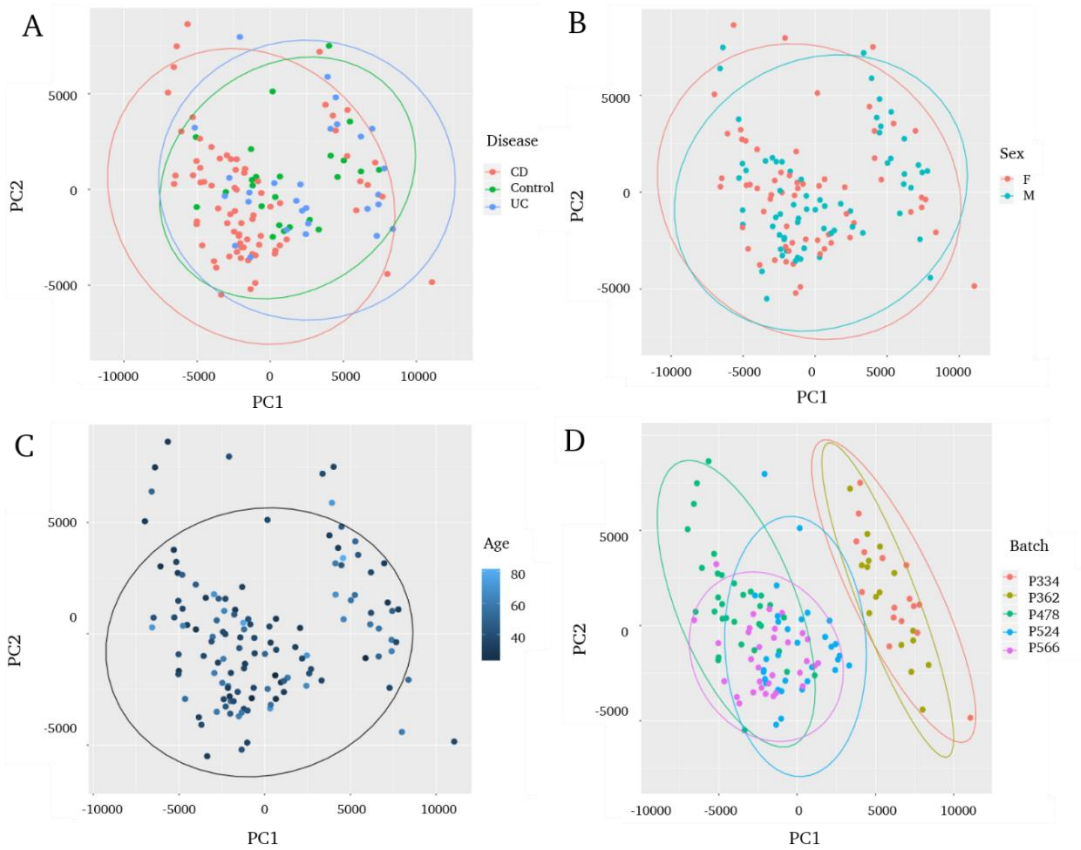


Figure 3.6 | Principle components analysis PC1 vs PC2
Principle component analysis (PCA) based on normalised gene expression per sample for the first (PC1) and second (PC2) principle component. The effects of disease status **(A)**, sex **(B)**, age **(C)** and batch **(D)** on PC1 vs PC2 distribution was investigated. The ellipse in plot A-D represents the 95% confidence interval for each subgroup. **(A)** Each dot represents a sample with red indicating a CD and blue indicating a UC diagnosis and green dots indicate control samples. **(B)** Samples coloured based on sex, where red dots are female and blue dots are male. **(C)** Samples coloured on age, from youngest (dark blue) through to oldest (light blue). **(D)** Sample coloured according to sequencing batch.

Random distribution was confirmed for disease status, sex, age and batch for all other principle components (PC3-PC6), with the exception of one major outlier observed in PC5 vs PC6 (**Figure 3.7A-D**). This sample, a female CD patient sequenced in batch P334, was therefore subsequently removed from downstream analysis.

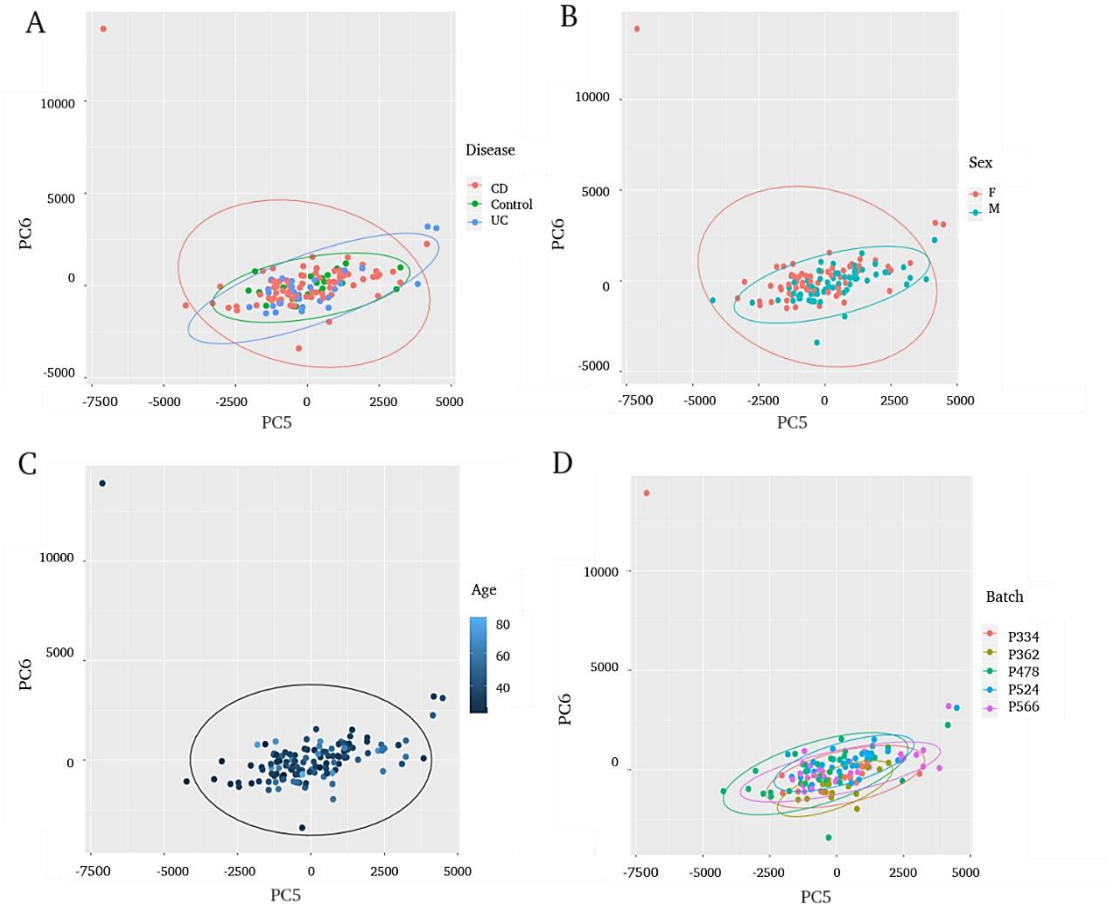


Figure 3.7 | Principle components analysis for PC5 vs PC6
Principle component analysis (PCA) based on normalised gene expression per sample for the fifth (PC5) and sixth (PC6) principle component. The effects of disease status **(A)**, sex **(B)**, age **(C)** and batch **(D)** on PC5 vs PC6 distribution was investigated. The ellipses in plots A-D represent the 95% confidence interval for each subgroup cluster. **(A)** Each dot represents a sample with red indicating CD, blue indicating UC and green indicating control samples. **(B)** Samples coloured with respect to gender, red dots are female and blue dots are male. **(C)** Samples coloured with respect to age, from youngest (dark blue) through to oldest (light blue). **(D)** Samples coloured according to sequencing batch.

3.6 Correlation between RNAseq datasets

To validate batch-effect corrections for the two library chemistries was sufficient (Illumina vs Epicentre), the correlation between them was investigated. Fold change values between IBD cases and controls, generated by the differential expression analysis, were used to perform a Pearson correlation test (**Figure 3.8**).

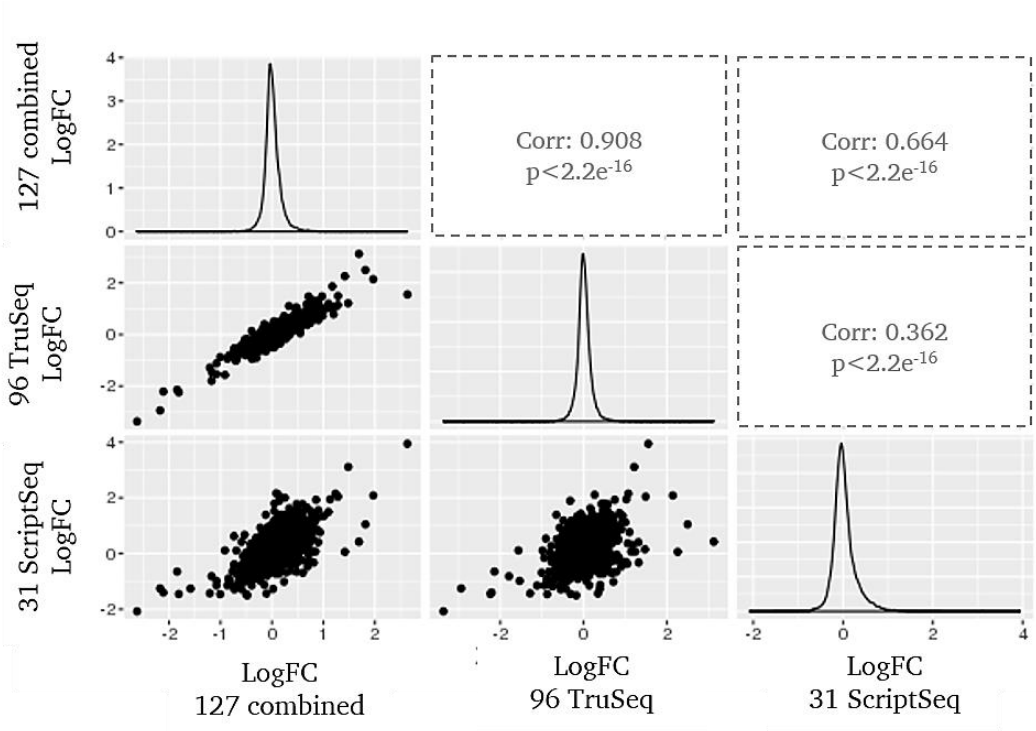


Figure 3.8 | Correlation test in RNAseq datasets
Pearson correlation test investigating correlation between fold change (FC) values of all transcripts (17,936) in the RNAseq datasets; 31 libraries generated with Epicentre ScriptSeq, 96 libraries generated with Illumia TruSeq and 127 samples from both chemistries combined. Each black dot represents a transcript within the scatterplots with a correlation and p-value related to each scatterplot shown top right. The histogram represents the normal distribution of each dataset.

Statistically significant positive correlation was observed between all pairwise comparisons ($p<2.2x10^{-16}$). These results show that the observed batch effect has been successfully corrected, and the combined dataset of 127 samples could be confidently used for further analyses including the differential expression analysis.

3.7 Discussion

Quality and reliability was assessed at every stage from RNA input to the final RNA sequencing data, confirming high quality data was generated. Due to circumstances two different chemistries to generate the whole RNA libraries were used. This resulted in a strong batch effects, by applying stringent QC measures and incorporating batch effect as a covariate into downstream analysis models, we successfully corrected for this batch effect. The combined dataset of 127 samples can be confidently used for further analyses including the differential expression analysis.

4. Qualitative and quantitative analysis of the transcriptome in the colon

The introduction of GWAS in 2006 initiated a major advance in unravelling the complex genetics of CD and UC. To date 27 GWA studies have been performed on CD and 21 on UC identifying > 200 IBD susceptibility loci [106], the highest for any single disease. Studies by Jostins et al, Mokry et al and Huang et al, all contribute to the theory that the majority of IBD associated SNPs are correlated with non-coding variants that perturb regulation of gene expression instead of directly altering gene function, making identification of causal genes challenging [92,112,149]. Recent progress in prioritising causal genes has been made through fine-mapping and expression quantitative trait (eQTL) analysis, with 16 of the known IBD susceptibility loci having been reduced to 1 causal variant at >95% probability [108,112]. A further 445 genes have been prioritised to be involved in IBD pathogenesis based on investigation of gene co-localisation, protein-protein interactions, functional connectivity of a gene within literature and eQTL analysis [107,150]. In order to attempt to understand how the majority of common susceptibility loci may influence IBD pathogenesis, the focus of research will have to shift focus from GWA studies to gene expression studies. Here whole RNA sequencing data from uninflamed large intestinal tissue was utilized to investigate the intestinal transcriptome including known IBD susceptibility loci. Differential gene expression was investigated between CD cases and controls, IBD vs controls, UC vs control and UC vs CD to further the understanding of CD and IBD pathogenesis.

4.1 Sample collection

The intestinal biopsies were taken from un-inflamed large intestine throughout the transverse and descending colon. Tissue samples from 75 CD patients, 28 UC patients and 24 controls were selected for whole RNA sequencing. The imbalance in cohort sizes was partly due to the original study design; previously the focus lay on CD patients only. Furthermore, the availability of patients played a role; there is a higher percentage of CD patients *versus* UC and non-

IBD patients having colonoscopies. It was decided to collect un-inflamed colonic tissue over inflamed tissue. The use of un-inflamed vs inflamed tissue within expression and functional studies is topic of discussion. Inflamed tissue samples show stronger expression signals of immunological and/or pro-inflammatory IBD-associated genes, but it is hard to distinguish if these genes are upregulated due to the primary cause of disease or just secondary to the inflammation that results. Using un-inflamed intestinal tissue will slightly reduce strength of the signal but will allow the separation between primary and secondary effects. An almost equal spread between males and females was observed; the CD cohort contained slightly more males (54% vs 46%), where the UC and control cohorts included slightly more females (46% and 42% *versus* 54% and 58%) (**Table 4.1**). The mean age within the cohorts was approximately 40 years of age, with UC patients being slightly older at mean 44.1 years (**Table 4.1**). Treatment regimens, from patients at the time of the biopsy being taken, were not recorded.

Table 4.1 | Patient demographics

	CD patients	UC patients	Controls (non-IBD)
Mean age	38.5	44.1	41.4
# Female	35	15	14
percentage	46%	54%	58%
# Male	41	13	10
percentage	54%	46%	42%

4.2 Qualitative analysis of the transcriptome in colon

Colonic transcriptome data, corresponding to expression levels (FPKM, Fragment per Kilobase per Million) for 56,260 known transcripts, was generated for 127 (24 controls, 28 UC and 75 CD) samples. FPKM quantifies expression based on read counts normalised against gene size. A threshold value of FPKM = 1 to distinguish true expression from background was used. Out of the 56,260 aligned transcripts 17,936 exhibited expression of FPKM ≥1, corresponding to 32% of the human transcriptome. 77% of these transcripts

corresponded to coding genes and approximately 3% were identified as long noncoding RNAs (lncRNAs). Furthermore, 30 highly expressed ribosomal RNAs were observed, indicating imperfect ribosomal RNA removal (ribodepletion) prior to sequencing. Consequently, all rRNA encoding transcripts were removed from further analyses.

To date, 224 IBD susceptibility loci have been identified (**Appendix 5**). In this study it was established that a total of 2,971 transcripts mapped to within 500 kb of one of the known IBD susceptibility loci. All IBD susceptibility loci were observed to contained a minimum of one transcript ≥1 FPKM, with the exception of IBD locus 6.03 (Chr6:14211961-15234463). Although none are expressed above background, locus 6.03 has been shown to contains 9 transcripts (all non-coding RNAs), none of which have previously been implicated in IBD pathogenesis. Within the 224 known IBD susceptibility loci, 16 have been reduced to a single causal variant and various others have been reduced to highly suggestive variants. Table 4.2 contains the genes directly affected or most likely affected by these identified causal and suggestive variants (**Table 4.2**). When investigated, the genes affected by the 16 causal variants all showed to exhibited expression above background (**Table 4.2**). Genes affected by the highly suggested variants showed expression levels to range from 1.9 to 83 FPKM; with *IRGM*, *IL12B* and *BTNL2* failing to reach expression levels above background (**Table 4.2**).

Table 4.2 Level of expression of top IBD susceptibility genes

Ensemble ID	Gene Name	Mean FPKM	IBD loci
ENSG00000162594	*IL23R*	4.4	1.07
ENSG00000134242	*PTPN22*	8	1.12
ENSG00000158714	*SLAMF8*	4.6	1.16
ENSG00000115267	*IFIH1*	14.6	2.11
ENSG00000115232	*ITGA4*	31.3	2.12
ENSG00000085978	*ATG16L1*	28.4	2.21
ENSG00000178623	*GPR35*	24.4	2.22
ENSG00000237693	*IRGM*	0.2	5.13
ENSG00000113302	*IL12B*	0.1	5.14
ENSG00000204290	*BTNL2*	0.2	6.08
ENSG00000019485	*PRDM1*	2.1	6.11

ENSG00000185811	IKZF1	10.5	7.08
ENSG00000096968	*JAK2*	20	9.01
ENSG00000187796	*CARD9*	1.3	9.04
ENSG00000134460	*IL2RA*	3.4	10.01
ENSG00000119919	*NKX2-3*	2.5	10.11
ENSG00000188906	*LRRK2*	3	12.03
ENSG00000166949	*SMAD3*	19.7	15.03
ENSG00000005844	*ITGAL*	11.5	16.04
ENSG00000167207	*NOD2*	1.4	16.05
ENSG00000121281	*ADCY7*	9.2	16.05
ENSG00000197943	*PLCG2*	15.7	16.07
ENSG00000175354	*PTPN2*	27.4	18.01
ENSG00000090339	*ICAM1*	6.7	19.02
ENSG00000101076	*HNF4A*	83.4	20.04
ENSG00000026036	*RTEL1-TNFRSF6B*	1.9	20.08

Each chromosome and susceptibility locus was investigated in more detail following the differential expression analysis (see Chapter 4.4).

4.3 Differential expression analysis in IBD

Differential expression (DE) analysis of all measurable transcripts (17,936) was performed using EdgeR. EdgeR implements a statistical methodology based on negative binomial distribution to assess differences in expression between subsets. The differential expression analysis was executed using count values per transcript normalised against library size. Furthermore, ERRC spike-in control data, age, sex, batch and disease type were included as covariates in the analysis. Reported p-values were adjusted for multiple testing by applying the Benjamini-Hochberg method, resulting in corrected false discovery rates (FDR). DE analysis was performed between CD cases and controls, IBD vs controls, UC vs controls and UC vs CD.

4.3.1 CD *versus* controls

Due to the imbalance in the patient cohorts: 75 CD patients, 28 UC patients and 24 controls, the CD versus control analysis carries the most power.

Differential expression (DE) analysis of all transcripts (17,936) between CD cases (n=75) and controls (n=24) was performed to identify transcripts of potential importance in CD pathogenesis (**Figure 4.1**).

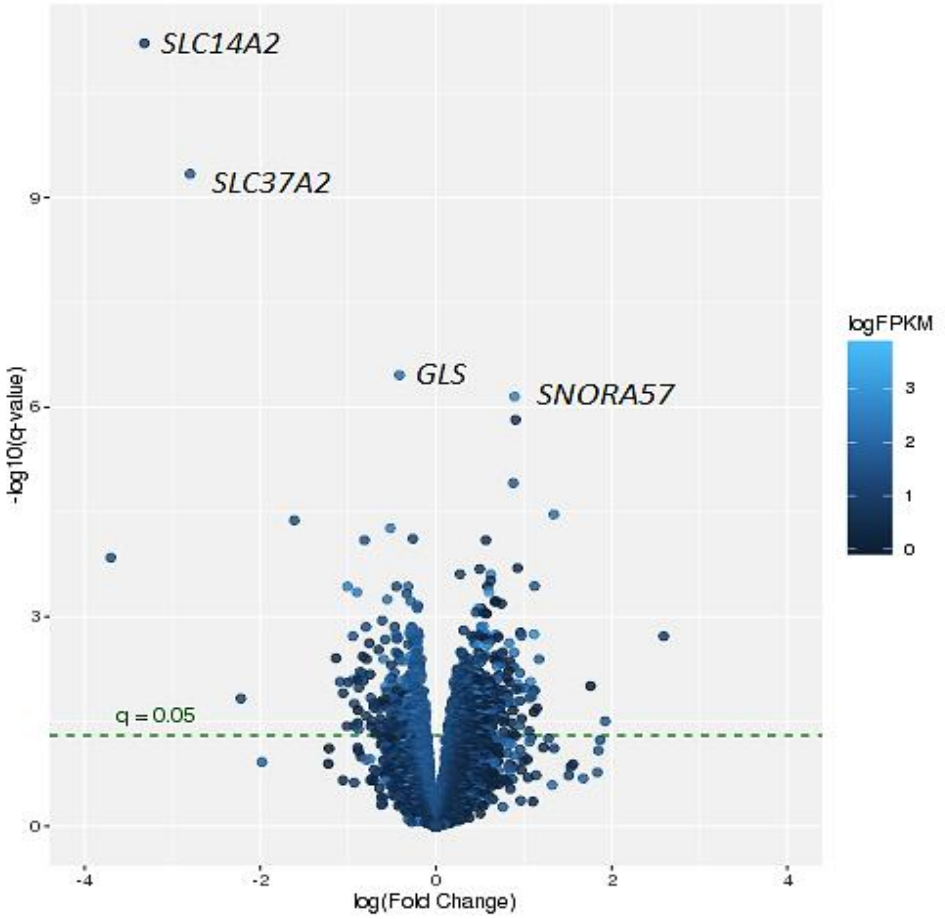


Figure 4.1 | DE analysis between CD and controls
Log fold change between CD cases and controls –log10(q-value) for DE analysis. Each dot represents a transcript with the colour indicating level of expression (logFPKM). Threshold for significance at q=0.05 (green line).

Evidence of significant difference in expression was found ($q \leq 0.05$) for 1,051 transcripts. The differentially expressed gene with greatest significance was *SLC14A2* (Solute Carrier Family 14 Member 2) ($q=6.18 \times 10^{-12}$), a member of the solute carriers (SLCs) family (**Figure 4.1**). *SLC14A2* was observed to be significantly downregulated in CD patients which is in accordance with previously reported findings [151]. Additionally, significant decreased CD expression of *SLC37A2* ($q=4.58 \times 10^{-10}$), another solute carrier gene, was

observed. SLCs can be subdivided into 47 families with more than 300 members, a subset of SLCs play a role in epithelial permeability and barrier function in the intestine and have previously been implicated in IBD development [152,153]. SLC14A2 has been suggested to be involved in the regulation of Urea flux into the gastrointestinal tract with higher Urea levels enhancing bacterial growth in the gut. Other top hits were *GLS* (Glutaminase) ($q=3.48x10^{-7}$) and *SNORA57* (Small Nucleolar RNA, H/ACA Box 57) ($q=1.86x10^{-7}$).

To facilitate prioritisation of DE transcripts based on their likely role in CD pathogenesis the significant findings were mapped to within 500 KB of the 224 known IBD susceptibility loci. When mapping the 1,051 identified significant CD (DE) transcripts to their genomic location, it was observed that 178 of these were located within 500kb of a known IBD susceptibility locus (**see appendix 5A**). When extending these boundaries to 1Mb on either side, this number increases to 289 (28% of DE genes), the 1Mb boundaries correlate to the boundaries applied in the expression quantitative trait (eQTL) analysis (see Chapter 6). The top hit was *GLS* (Glutaminase), which encodes an enzyme involved in the hydrolysis of glutamine into glutamate and ammonia. This gene shows reduced expression in CD cases versus healthy controls. Glutamine is an amino acid required for protein biosynthesis and an important energy source for a wide variety of cells, including rapid dividing immune cells and gut mucosal cells which are two major consumers of Glutamine metabolism [154]. It has been shown that reduced Glutamine metabolism could lead to reduced gut mucosal integrity and increased gut permeability to allergens and pathogens, causing intestinal inflammation [154,155]. Additionally, transcripts included *TRAF3IP2* (Nuclear Factor NK-Kappa-B activator), *DENND1B* (DENN Domain Containing 1B) and *TNFRSF14* (Tumour Necrosis Factor Receptor Superfamily Member 14). *TRAF3IP2* and *DENND1B* were identified to show lower expression in CD cases whereas, *TNFRSF14* showed increased expression. TRAF3IP2 (or ACT1) has been reported to activate transcription factor NF-$_k$B through I_kB kinase (IKK) activation as well as activate Jun kinase (JNK). NK-$_k$B is known to play a vital role in immune and inflammatory responses, cell

survival and stress responses [156]. DENND1B has been shown to regulate T-cell receptor (TCR) internalization in Th2 cells, and when disrupted TCR signalling was enhanced leading to increased cytokine production and immune responses [157]. TNFRSF14 (or HVEM) has been reported to exhibit ligand dependant bi-directional signalling. TNFRSF14 has been suggested to induce NF-$_k$B activation, triggering pro-inflammatory and cell survival genes; however it also mediates inhibitory signalling [158]. It has been proposed that TNFRSF14 mediated signalling represents an important immune regulator, specifically in mucosal surfaces, in autoimmunity and infection [159].

4.3.2 IBD *versus* controls

By combining the CD and UC patients, transcripts contributing to general IBD pathogenies were investigated. IBD specific colonic transcript expression was examined by performing DE analysis for IBD (n=104) vs controls (n=24) (**Figure 4.2**).

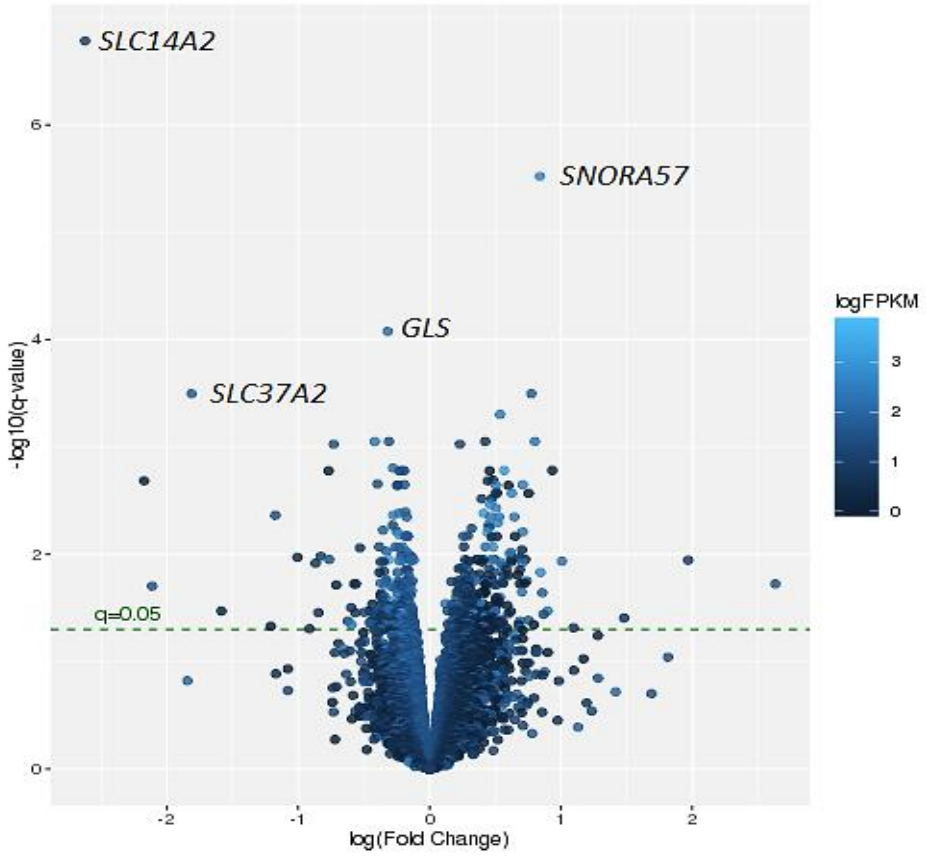


Figure 4.2 | DE analysis between IBD cases and controls
Log fold change between IBD cases and controls vs –log10(q-value) for DE analysis. Each dot represents a transcript with the colour indicating level of expression (logFPKM). Threshold for significance at q=0.05 (green line).

526 transcripts were identified with a significant difference in expression ($q<0.05$). The direction of effect was observed to be approximately equally distributed with 250 transcripts showing a decreased fold change in IBD and 276 transcripts showing and increased fold change. The top hits *SLC14A2*, *SLC37A2*, *GLS* and *SNORA57* corresponding to the top hits found in the CD vs controls analysis. However, the magnitude of the significance has decreased in all four top hits, most notably *SLC14A2* with $q = 6.18\text{x}10^{-12}$ to $q = 1.65\text{x}10^{-7}$ and *GLS* $q = 3.48\text{x}10^{-7}$ to $q = 8.4\text{x}10^{-5}$, suggesting the effect seen in the IBD cases vs controls was driven by the CD cases.

When mapping the 526 identified significant IBD (DE) transcripts to their genomic location, it was observed that 80 transcripts with FPKM ≥ 1 and $q \leq 0.05$ were located within 500kb of a known IBD susceptibility locus (**see**

appendix 5B), this number increased to 127 (25%) when extending the boundaries to 1Mb. An overlap of 59 DE genes was identified in both DE analyses, with 21 transcripts being solely differentially expressed in the IBD vs controls analysis. Top hit genes including *GLS* (Glutaminase), *HLA-DRB5* (major histocompatibility complex, class II, DR Beta 5), *VIL1* (Villin 1) and *ITCH* (Itchy E3 Ubiquitin Protein Ligase) were observed to have significantly reduced expression in IBD cases. *HLA-DRB5,* a HLA class II molecule, is known to be expressed in antigen presenting cells and involved in regulation of immune responses. VIL1 plays a role in intestinal cell morphology and cell migration, and over expression of VIL1 has been shown to protect against apoptosis of intestinal epithelium cells [160-162]. ITCH has been shown to be a component of an ubiquitin-editing protein complex involved in the control of inflammatory signalling pathways, most notably, ITCH ubiquitinated RIP2 to allows differential NOD2:RIP2 signalling [163]. Furthermore, 3 non-coding RNAs; *RNY1*, *SNORA74B* and *SNORA42*, were identified.

4.3.3 UC *versus* controls

DE analysis for UC cases (n=28) vs controls (n=24) was performed to identify UC specific transcripts (**Figure 4.3**).

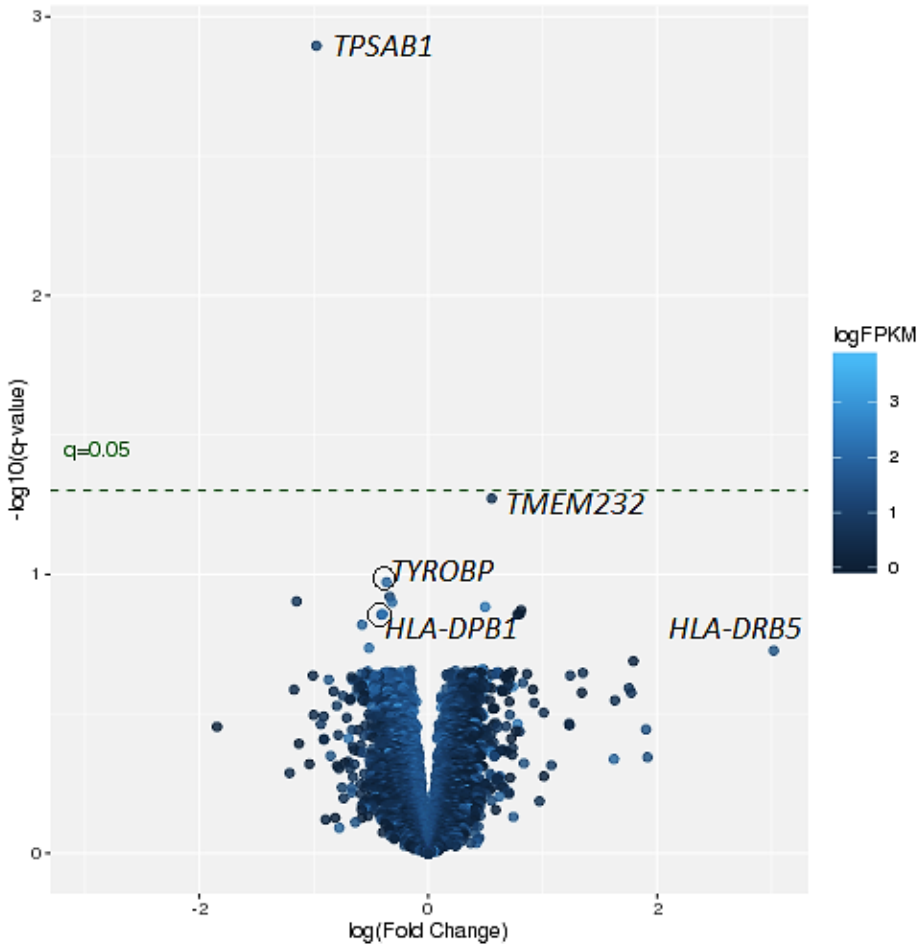


Figure 4.3 | DE analysis between UC and controls
Log fold change between UC cases and controls vs –log10(q-value) for DE analysis. Each dot represents a transcript with the colour indicating level of expression (logFPKM). Threshold for significance at q=0.05 (green line).

A plateau formation was observed around a q-value of approximately 0.2 (-log10 = 0.7) (**Figure 4.3**), this was contributed to the reduced power within the UC (n = 28) vs controls (n = 24) analysis. Only one gene, *TPSAB1* (Tryptase Alpha/Beta 1), reached significance while *TMEM232* (transmembrane protein 232) came close to significance with q=0.053. A reduced expression of *TPSAB1* within UC cases was observed whereas, *TMEM232* expression was increased. *TPSAB1* encoding Tryptase, a protein secreted by mast cells, has been implicated in weakening the inflammatory response of β tryptase [164]. Other transcripts located above the plateau, approximately –log10(1), show relatively high expression levels (FPKM) and included *HLA-DPB1* (major histocompatibility complex, class II, DP Beta 1),

TYROBP (TYRO protein tyrosine kinase binding protein) and *HLA-DBR5*. TUROBP (or DAP12) a transmembrane protein has been suggested to associate with killer-cell inhibitory receptors (KIR) in natural killer (NK) cells, regulating cell activation. Both HLA-DPB1 and HLA-DBR5 have been identified as HLA class II molecules involved in antigen presentation in response to extracellular proteins. Interestingly, the direction of effect in DE observed for *HLA-DPB1* and *HLA-DBR5* was opposing; exhibited fold change was negative for *HLA-DPB1* and positive for *HLA-DBR5*. In addition, when mapping the DE transcripts to the known IBD susceptibility genomic locations *HLA-DPB1* and *HLA-DBR5* were observed to have the lowest q-value of transcripts within an IBD locus, although not significant. Taking into consideration the small sample size – n = 24 *versus* n = 28 - these results suggested the DE analysis was underpowered and an increase in sample size should be considered in the future.

4.3.4 UC *versus* CD

When comparing UC (n=24) versus CD (n=75) to identify disease specific genes, 696 transcripts were identified as significantly differentially expressed (**Figure 4.4**).

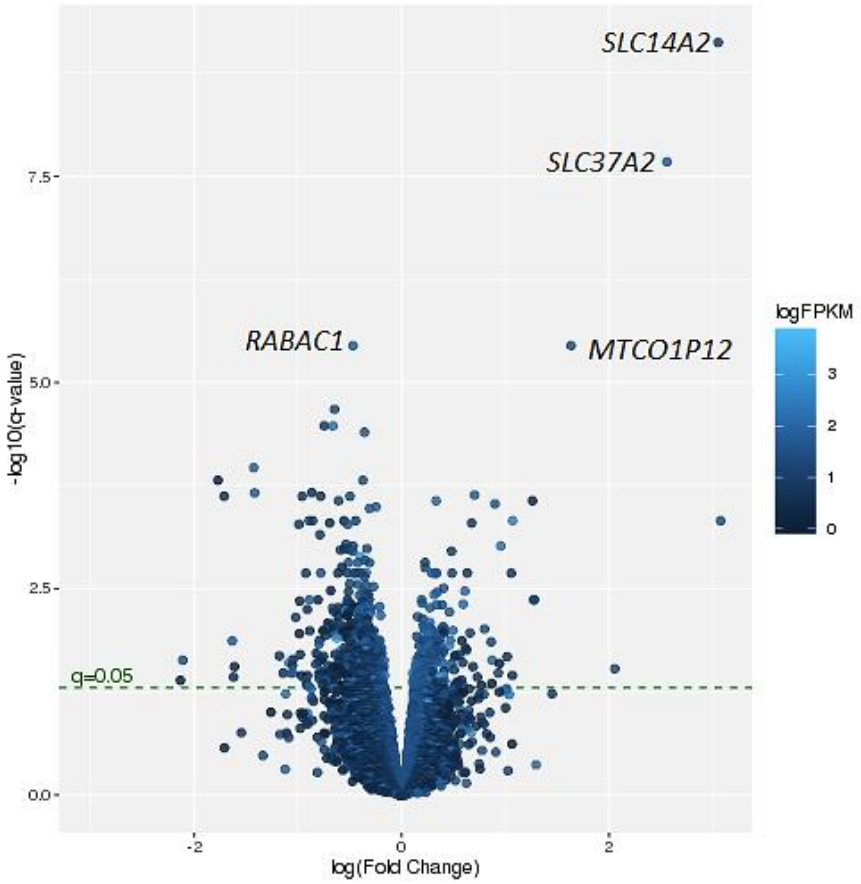


Figure 4.4 | DE analysis between UC vs CD
Log fold change between UC and CD vs –log10(q-value) for DE analysis. Each dot represents a transcript with the colour indicating level of expression (logFPKM). Threshold for significance at q=0.05 (green line).

It was observed that a greater proportion of transcripts was more highly expressed in CD cases compared to UC (484, 70%), where the remaining 212 transcripts (30%) demonstrated increased expression within UC cases. The top hits for both negative and positive fold change values were previously identified in the CD vs controls analysis indicating that the majority of the observed effect was driven by the CD cases, or that UC sample size was underpowered. *SLC14A2*, *SLC37A2* and *MTCO1P12* (MT-CO1 Pseudogene 12) exhibited higher expression in UC compared to CD. *SLC14A2* and *SLC37A2* are members of the solute carrier family, and were previously identified in both the IBD and CD vs control analyses. *MTCO1P12* is a pseudogene affiliated with the lncRNA class. *RABAC1* (Rab Acceptor 1), *C19orf60* and *SYTL1* (Synaptotagmin like 1)

exhibited higher expression within CD versus UC, equally they were identified as overexpressed within the CD vs control analysis.

Within the UC vs CD DE analysis, 134 DE transcripts were identified to be located within 500 kb and 213 (31%) within 1Mb of the 224 known IBD loci (**see Appendix 5C**). Of these, 49 were also identified in the CD vs control analysis and 7 were identified in the IBD vs control analysis. Top hits included, *TMEM259* (Transmembrane protein 259) ($q = 4.01x10^{-5}$), *GAL3ST2* (Galactose-3-O-Sulfotransferase 2) ($q = 2.39x10^{-4}$) and *GPC1* (Glypican 1) at $q = 4.72x10^{-4}$. *TMEM259* and *GAL3ST2* were uniquely identified within the UC vs CD comparison, both exhibited lower expression levels within UC vs CD. GAL3ST2 is thought to be responsible for sulfotransferase within human colonic mucins and upregulated in response to inflammatory stimuli. Appendix 5C contains all CD vs UC differentially expressed genes.

4.4 Prioritisation of potential causal genes in IBD

Multiple differential expression analyses were successfully performed on the generated RNAseq data set including CD, IBD, UC vs controls and UC vs CD (**Table 4.2**). The loci boundaries were extended by 500kB and 1Mb in an attempt to prioritise genes potentially casual in IBD. IBD susceptibility SNPs are known to be located in between coding regions and are speculated to affect regulatory elements or transcription factors of nearby genes. By extending the loci boundaries we aim to include genes located outside the IBD loci but whose expression is affected by IBD risk SNPs located within the IBD loci. When calculating the number of DE genes expected to fall within an IBD susceptibility loci by random chance; based on the combined size of the IBD susceptibility loci and size of the genome, it was observed that we identify a higher amount of genes to be differentially expressed than expected by chance. Based on chance alone we would expect to identify 9% of DE genes to be located within 500Kb and approximately 16% within 1Mb of an IBD susceptibility locus. In our analysis, of the total DE identified genes approximately 15-20% fell within the boundaries of known IBD susceptibility loci locations extended by 500 kb

on either side (**Table 4.3**). This percentage increased to 25-31% by extending the boundaries to 1Mb (**Table 4.3**).

Table 4.3 | Differentially expression analysis

Number of diff. expressed genes (q<0.05)	CD vs Control	IBD vs Control	UC vs Control	UC vs CD
Whole transcriptome	1051	526	1	696
Within 500Kb of IBD loci	178	80	0	134
Within 1Mb of IBD loci	289	127	0	213

In addition to their location in a known IBD susceptibility locus, overlap between the DE gene lists was assessed (**Figure 4.5**). CD vs control DE genes showed the highest overlap; 73% of genes identified in the IBD vs control analysis were also found in the CD vs control analysis and 36% of UC vs CD DE genes showed an overlap with CD vs control DE genes (**Figure 4.5**). This suggests that the CD cases are driving much of the observed overall results.

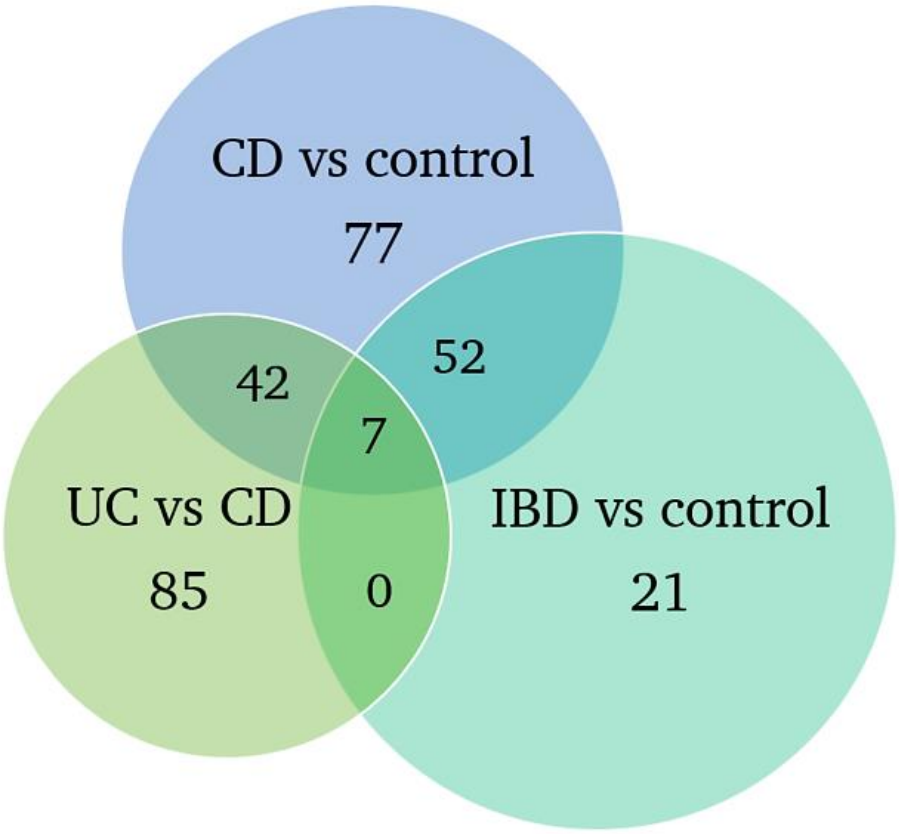


Figure 4.5 | Overlap differentially expression analysis results
Venn diagram showing the overlap between the differential expression analysis results between IBD vs control (n=21 unique), CD vs controls (n=77 unique) and UC vs CD (n=85 unique). Gene numbers based on genes located within 500Kb of an IBD susceptibility loci.

Overall, this analysis revealed a subset of transcripts potentially important to IBD pathogenesis based on their phenotype specific expression and genomic location.

4.4.1 Previously prioritised genes at known IBD loci

The number and size of known IBD susceptibility loci combined with the lack of definitive single causal variants in most loci [112], results in an immense amount of potential causal genes in IBD. In order to prioritise these genes Jostins *et al.* and Liu *et al.* performed DAPPLE, GRAIL and eQTL analyses [92,107]. DAPPLE to evaluate the disease association of genes via protein-protein interactions functional connectivity and correlation between disease associated SNPs and gene expression. More recently, Huang *et al.* and De Lange *et al.*, have performed a summary-statistic fine-mapping and eQTL analysis on these

loci in an attempt to prioritise genes [108,112]. By combining all these findings [92,107,108,112], 461 genes were prioritised to be the most likely candidates to be causal in IBD, covering 168 out of 224 IBD loci. For the remaining 56 IBD loci however, gene prioritisation has not been possible due to the lack of available functional information on the genes within them or the existence of multiple independent correlated association signals.

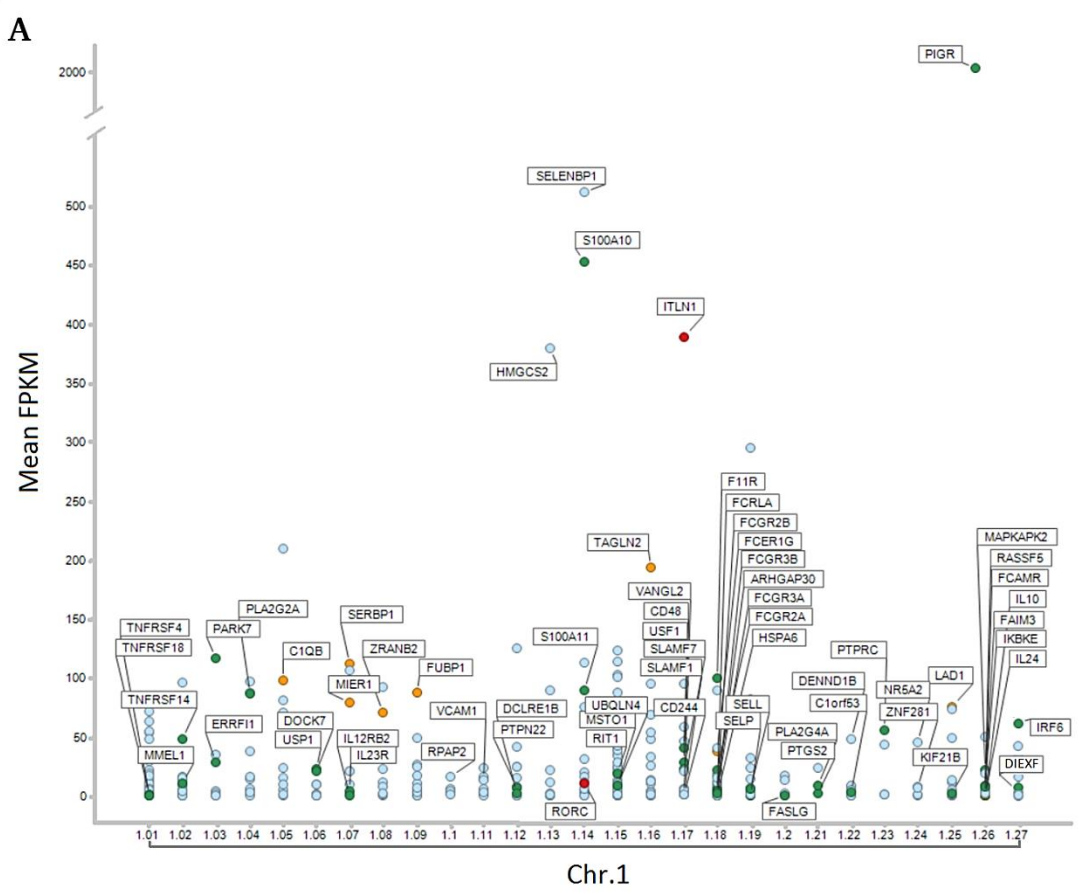
4.4.2 Validation of previously prioritised genes in IBD loci by differential expression analysis of colonic RNAseq data

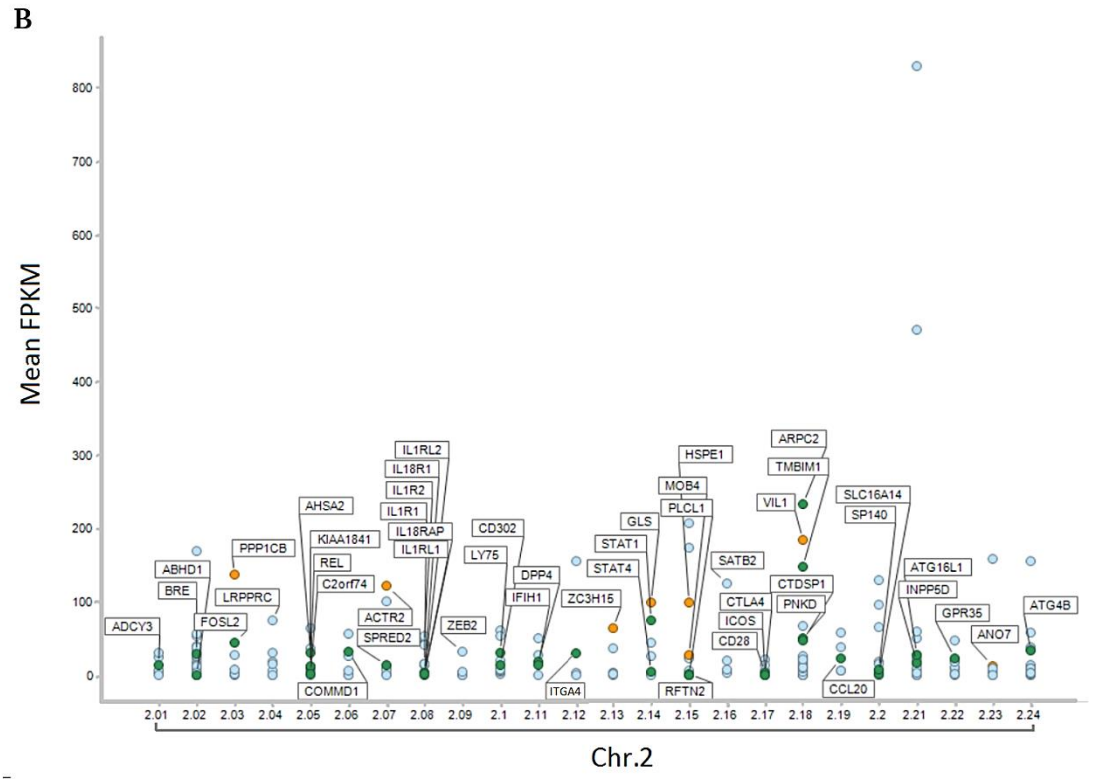
The results of the IBD vs control differential expression (DE) analyses were compared with the list of 461 genes previously prioritised to be causal. 8 genes were identified in both lists (**Table 4.4**, **Figure 4.6**).

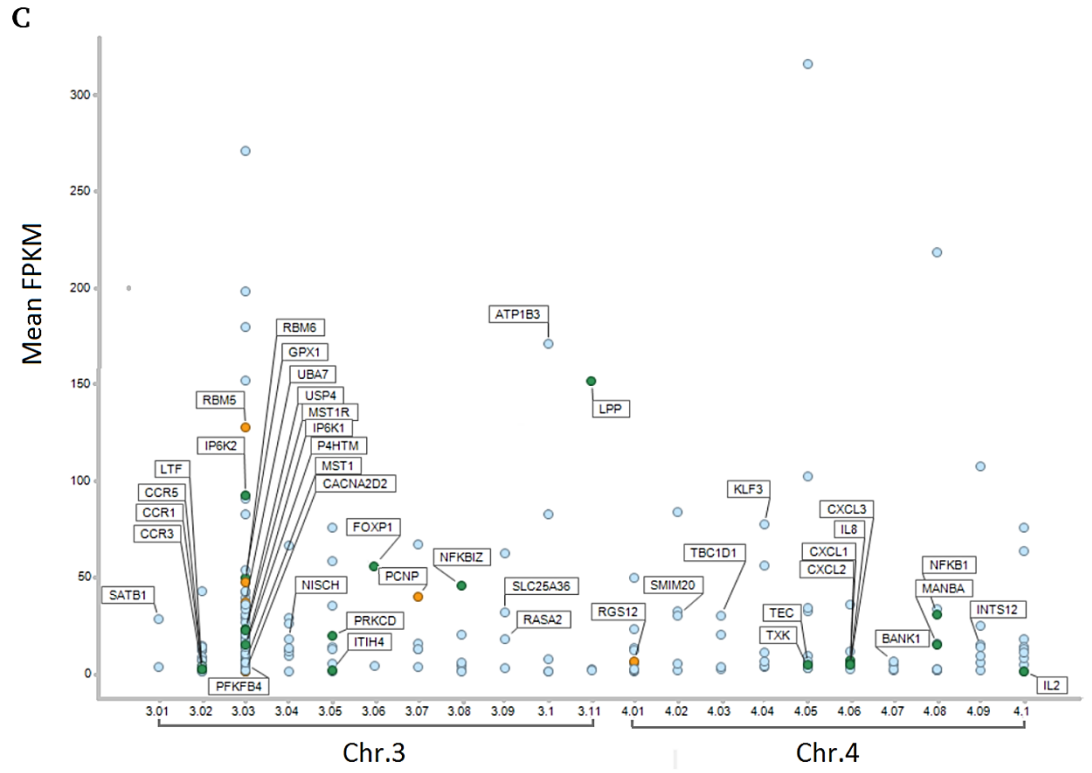
Table 4.4 | Prioritised and differently expressed genes

Gene Name	IBD Locus	locus location (Mbp)	q-value	Summary of function
RORC	1.14	151.2-152.3	$2.9x10^{-02}$	Key regulator of Th17 cell differentiation
CD244	1.17	160.3-161.4	$2.9x10^{-02}$	Mediates non-MHC restricted killing
ITLN1	1.17	160.3-161.4	$3.9x10^{-02}$	May be involved in the defence against microorganisms
PFKFB4	3.03	47.9-51.6	$4.2x10^{-02}$	Regulates of fructose-2,6-bisphophate concentrations
FAM49B	8.06	130.0-131.1	$2.4x10^{-02}$	Involved in antigen presentation
RAPGEF3	12.04	47.7-48.7	$4.2x10^{-02}$	Modulates cAMP-induced control of endothelial barrier function
CDH13	16.08	82.3-83.4	$1.9x10^{-02}$	Protects endothelial cell from apoptosis due to oxidative stress
CTSZ	20.07	57.3-58.3	$4.0x10^{-02}$	Lysosomal cysteine proteinase

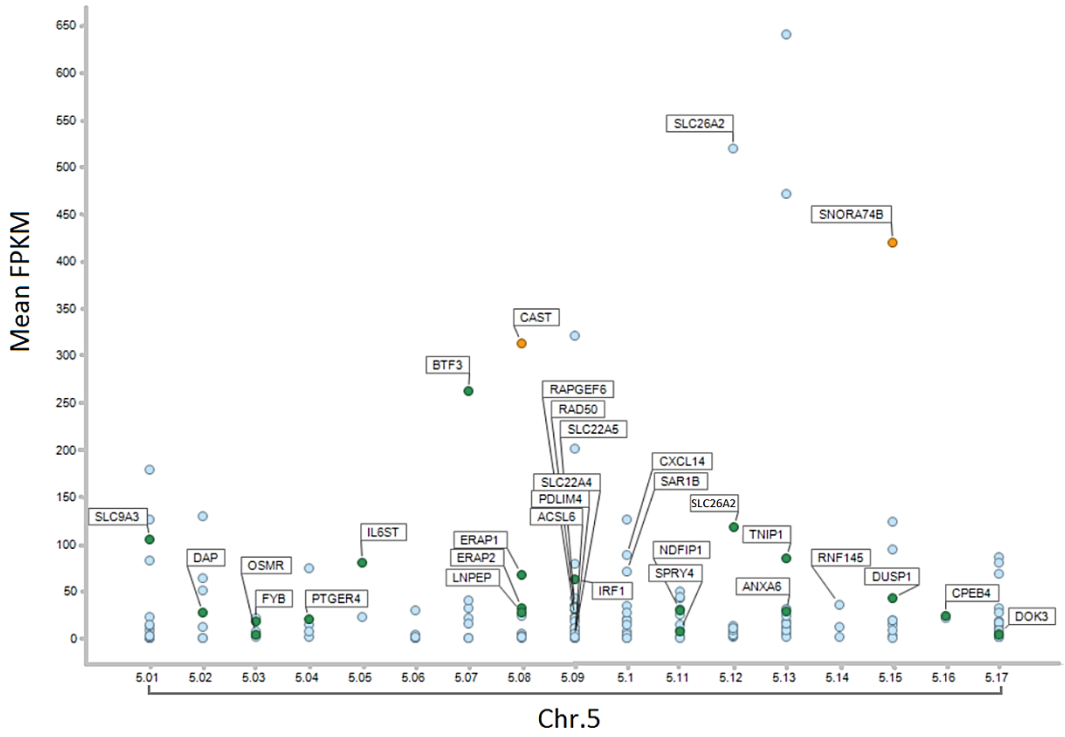
Moreover, 7 of the 8 genes are reported to be involved within immune responses or endothelial function both of which are known to be important in IBD (**Table 4.4)**. It was observed that 10 out of the 56 IBD loci failing to have previously prioritised genes contained differentially expressed genes. Furthermore, 46 IBD loci out of 224 did not contain either significant DE genes or prior prioritised genes which showed expression above background (≥ 1 FPKM), with one IBD locus (6.03) completely devoid of genes (**Figure 4.6**).



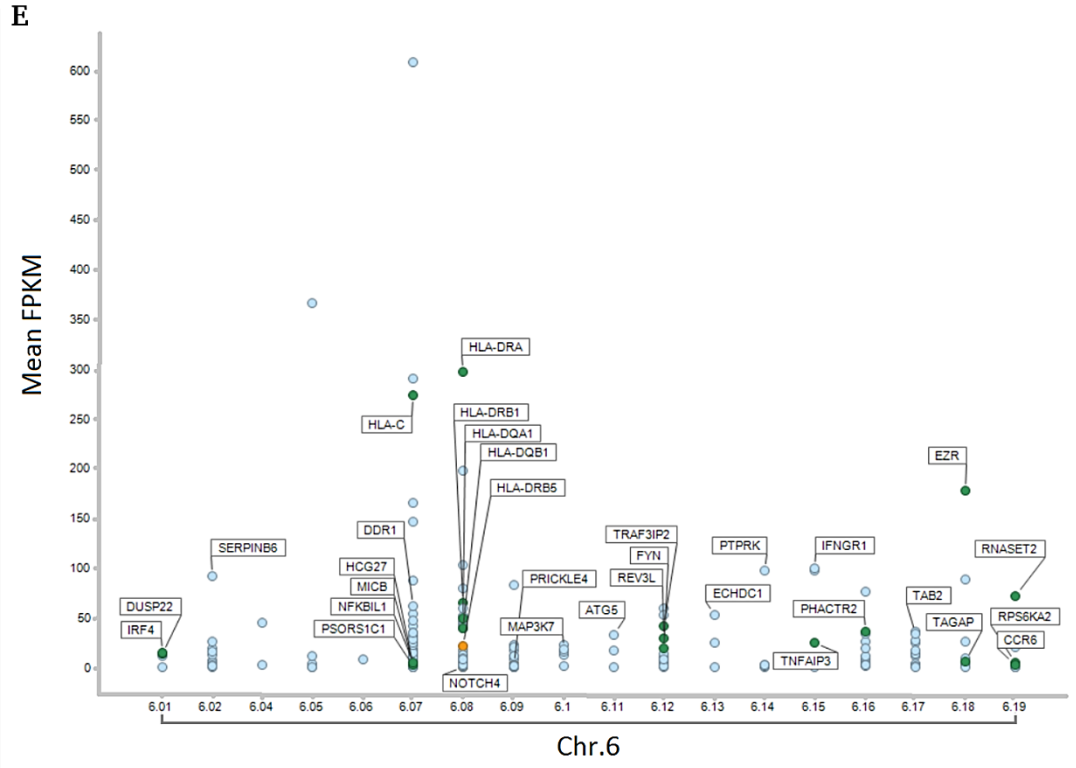

B
Mean FPKM
800
700
600
500
400
300
200
100
0
ADCY3
ABHD1
BRE
PPP1CB
LRPPRC
FOSL2
AHSA2
KIAA1841
REL
C2orf74
ACTR2
SPRED2
COMMD1
IL1RL2
IL18R1
IL1R2
IL1R1
IL18RAP
IL1RL1
ZEB2
CD302
LY75
DPP4
IFIH1
ITGA4
ZC3H15
STAT1
STAT4
GLS
HSPE1
MOB4
PLCL1
RFTN2
SATB2
CTLA4
ICOS
CD28
ARPC2
TMBIM1
VIL1
CTDSP1
PNKD
CCL20
SLC16A14
SP140
INPP5D
ATG16L1
GPR35
ANO7
ATG4B
2.01 2.02 2.03 2.04 2.05 2.06 2.07 2.08 2.09 2.1 2.11 2.12 2.13 2.14 2.15 2.16 2.17 2.18 2.19 2.2 2.21 2.22 2.23 2.24
Chr.2


C
Mean FPKM
300
250
200
150
100
50
0
SATB1
LTF
CCR5
CCR1
CCR3
RBM5
IP6K2
RBM6
GPX1
UBA7
USP4
MST1R
IP6K1
P4HTM
MST1
CACNA2D2
FOXP1
NISCH
PCNP
NFKBIZ
PRKCD
ITIH4
PFKFB4
ATP1B3
SLC25A36
RASA2
LPP
RGS12
SMIM20
TBC1D1
KLF3
TEC
TXK
CXCL3
IL8
CXCL1
CXCL2
BANK1
NFKB1
MANBA
INTS12
IL2
3.01 3.02 3.03 3.04 3.05 3.06 3.07 3.08 3.09 3.1 3.11 4.01 4.02 4.03 4.04 4.05 4.06 4.07 4.08 4.09 4.1
Chr.3
Chr.4

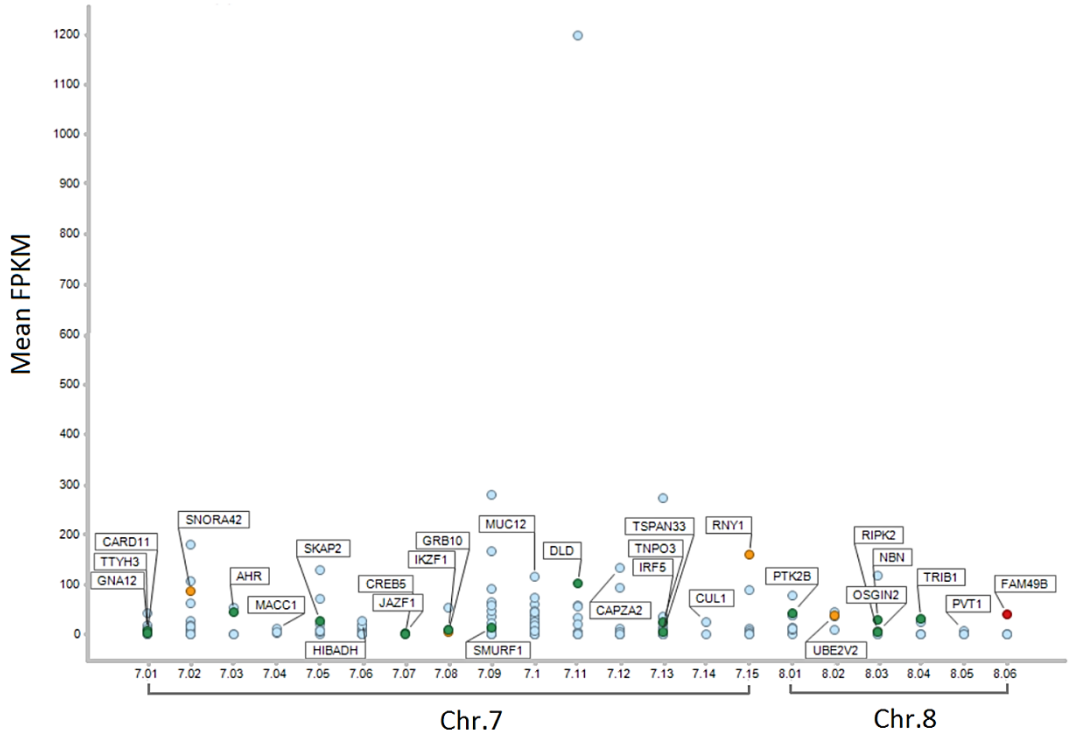
D


Mean FPKM
650
600
550
500
450
400
350
300
250
200
150
100
50
0
SLC26A2
SNORA74B
CAST
BTF3
RAPGEF6
RAD50
SLC22A5
CXCL14
SAR1B
SLC22A4
PDLIM4
ACSL6
SLC26A2
TNIP1
SLC9A3
IL6ST
ERAP1
ERAP2
LNPEP
NDFIP1
SPRY4
RNF145
DUSP1
DAP
OSMR
FYB
PTGER4
IRF1
ANXA6
CPEB4
DOK3
5.01
5.02
5.03
5.04
5.05
5.06
5.07
5.08
5.09
5.1
5.11
5.12
5.13
5.14
5.15
5.16
5.17
Chr.5

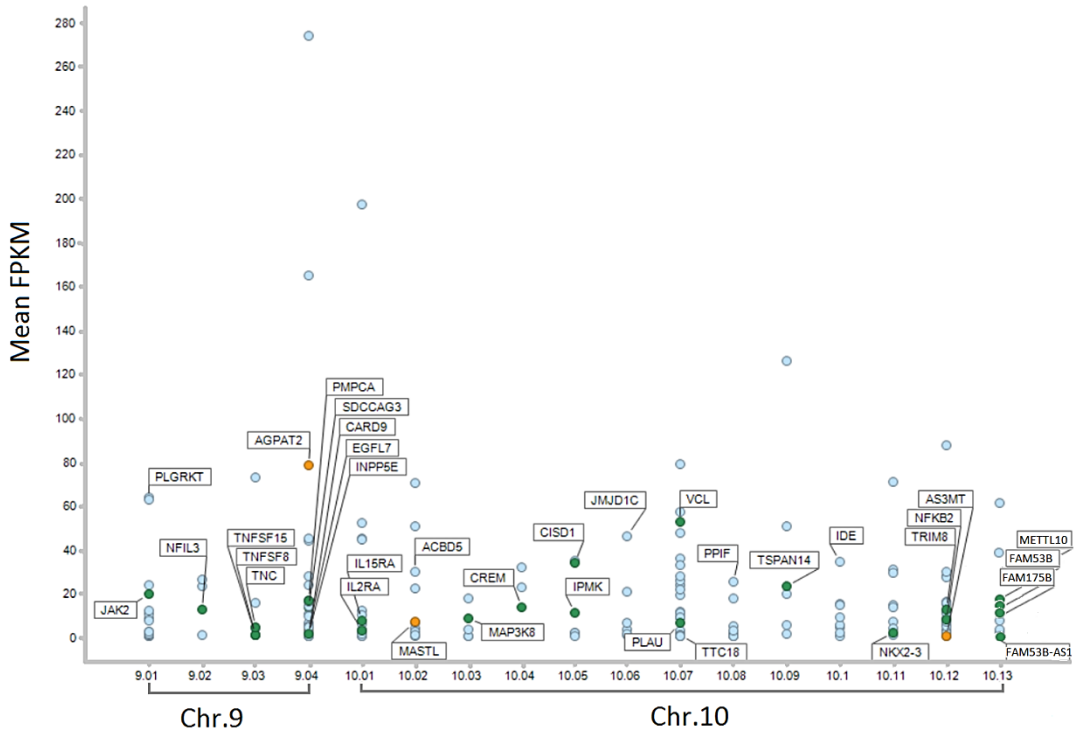
E


Mean FPKM
600
550
500
450
400
350
300
250
200
150
100
50
0
HLA-DRA
HLA-DRB1
HLA-C
HLA-DQA1
HLA-DQB1
HLA-DRB5
EZR
DDR1
TRAF3IP2
SERPINB6
FYN
PTPRK
IFNGR1
RNASET2
HCG27
PRICKLE4
REV3L
ECHDC1
MICB
TAB2
DUSP22
NFKBIL1
ATG5
PHACTR2
TAGAP
RPS6KA2
IRF4
PSORS1C1
MAP3K7
TNFAIP3
CCR6
NOTCH4
6.01
6.02
6.04
6.05
6.06
6.07
6.08
6.09
6.1
6.11
6.12
6.13
6.14
6.15
6.16
6.17
6.18
6.19
Chr.6

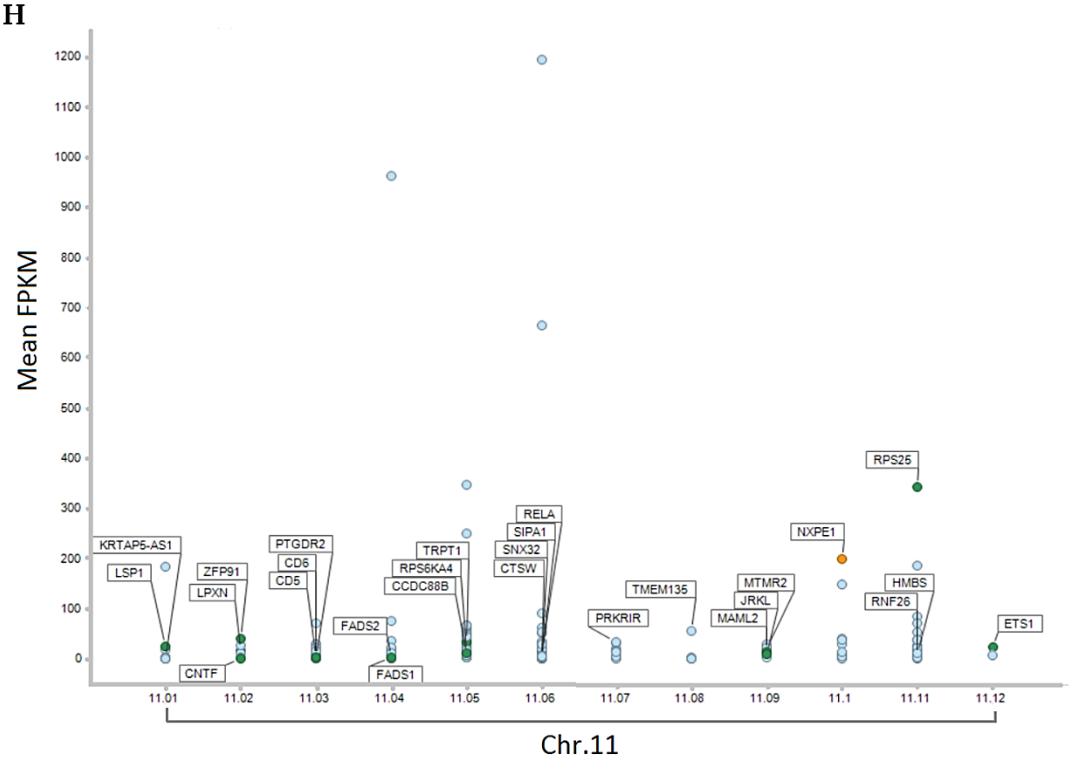
F


Mean FPKM
1200
1100
1000
900
800
700
600
500
400
300
200
100
0
CARD11
TTYH3
GNA12
SNORA42
AHR
MACC1
SKAP2
HIBADH
CREB5
JAZF1
GRB10
IKZF1
MUC12
SMURF1
DLD
CAPZA2
TSPAN33
TNPO3
IRF5
RNY1
CUL1
PTK2B
UBE2V2
RIPK2
NBN
OSGIN2
TRIB1
PVT1
FAM49B
7.01
7.02
7.03
7.04
7.05
7.06
7.07
7.08
7.09
7.1
7.11
7.12
7.13
7.14
7.15
8.01
8.02
8.03
8.04
8.05
8.06
Chr.7
Chr.8

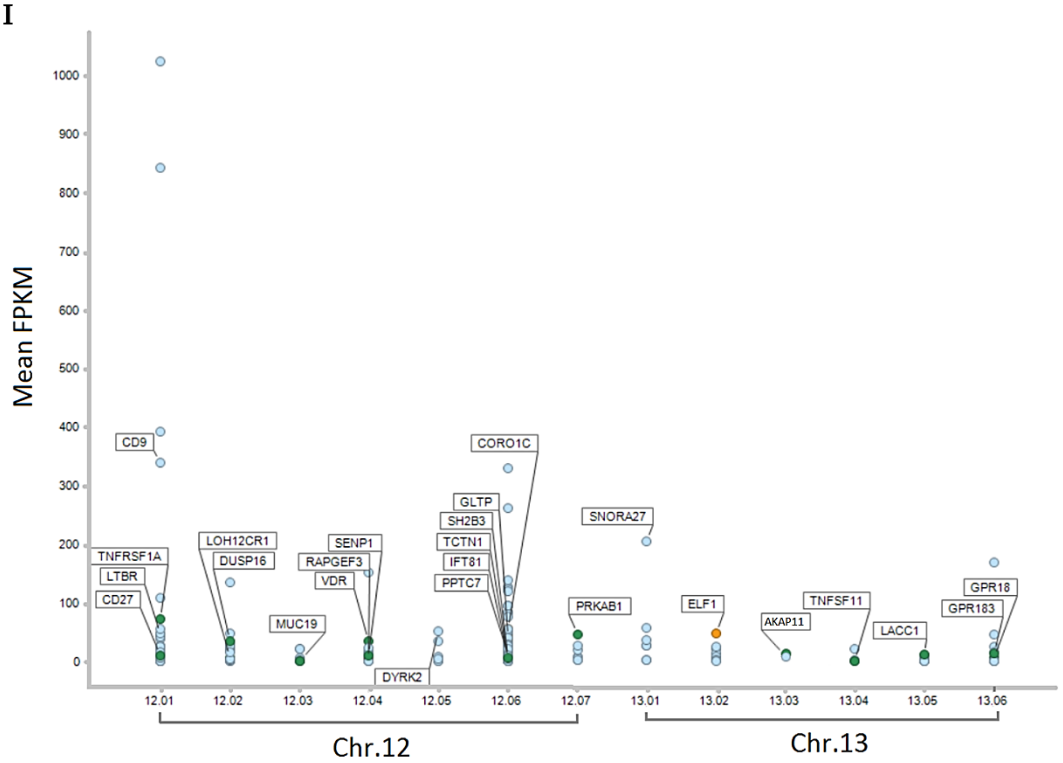
G

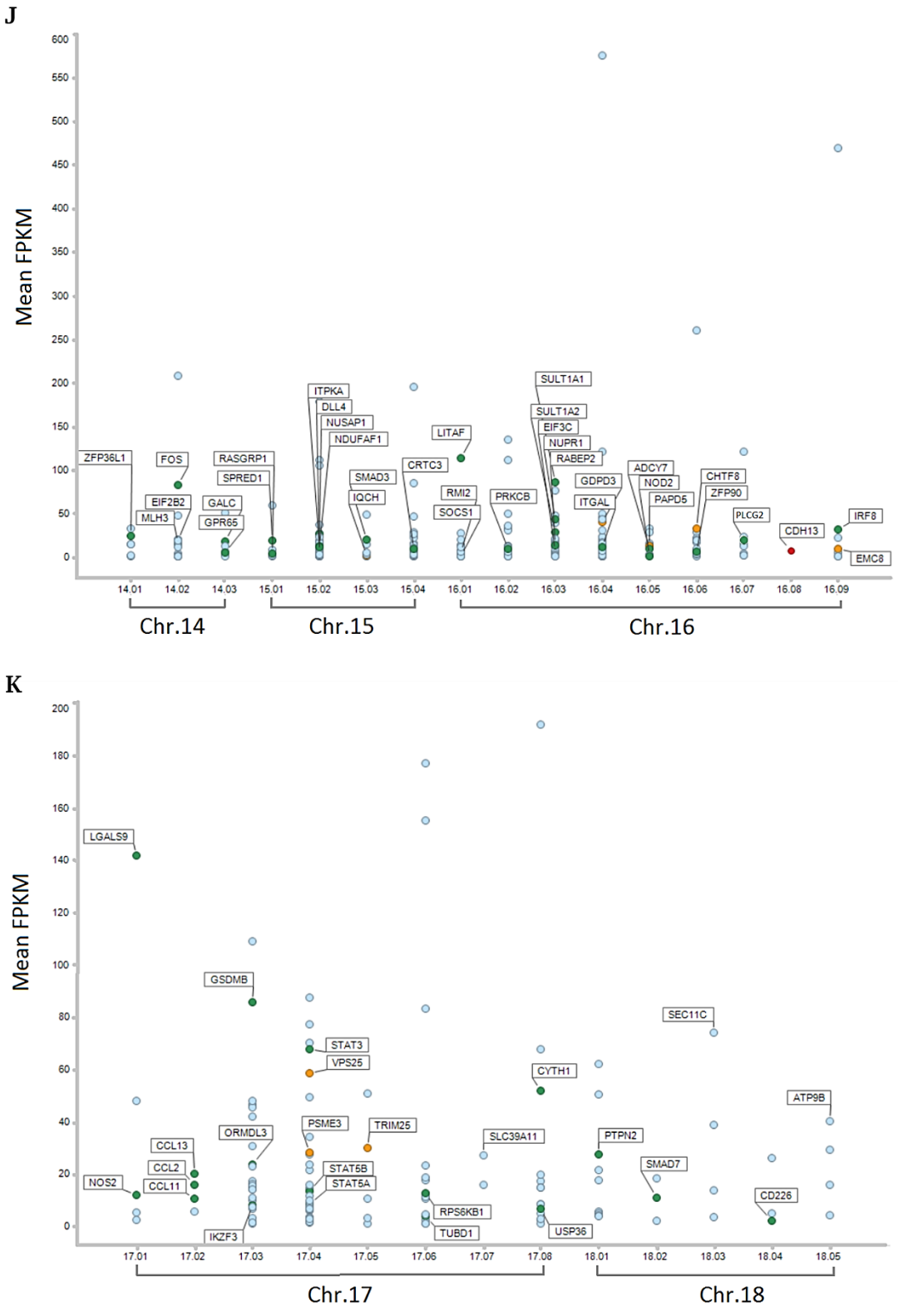

Mean FPKM
280
260
240
220
200
180
160
140
120
100
80
60
40
20
0
JAK2
PLGRKT
NFIL3
TNFSF15
TNFSF8
TNC
AGPAT2
PMPCA
SDCCAG3
CARD9
EGFL7
INPP5E
IL15RA
IL2RA
ACBD5
MASTL
CREM
MAP3K8
CISD1
IPMK
JMJD1C
PLAU
VCL
TTC18
PPIF
TSPAN14
IDE
AS3MT
NFKB2
TRIM8
NKX2-3
METTL10
FAM53B
FAM175B
FAM53B-AS1
9.01
9.02
9.03
9.04
10.01
10.02
10.03
10.04
10.05
10.06
10.07
10.08
10.09
10.1
10.11
10.12
10.13
Chr.9
Chr.10

H


Mean FPKM
1200
1100
1000
900
800
700
600
500
400
300
200
100
0
KRTAP5-AS1
LSP1
ZFP91
LPXN
CNTF
PTGDR2
CD6
CD5
FADS2
FADS1
TRPT1
RPS6KA4
CCDC88B
RELA
SIPA1
SNX32
CTSW
PRKRIR
TMEM135
MTMR2
JRKL
MAML2
NXPE1
RPS25
HMBS
RNF26
ETS1
11.01
11.02
11.03
11.04
11.05
11.06
11.07
11.08
11.09
11.1
11.11
11.12
Chr.11

I


Mean FPKM
1000
900
800
700
600
500
400
300
200
100
0
CD9
TNFRSF1A
LTBR
CD27
LOH12CR1
DUSP16
MUC19
SENP1
RAPGEF3
VDR
DYRK2
CORO1C
GLTP
SH2B3
TCTN1
IFT81
PPTC7
PRKAB1
SNORA27
ELF1
AKAP11
TNFSF11
LACC1
GPR18
GPR183
12.01
12.02
12.03
12.04
12.05
12.06
12.07
13.01
13.02
13.03
13.04
13.05
13.06
Chr.12
Chr.13


J
Mean FPKM
600
550
500
450
400
350
300
250
200
150
100
50
0
ZFP36L1
FOS
EIF2B2
MLH3
GALC
GPR65
RASGRP1
SPRED1
ITPKA
DLL4
NUSAP1
NDUFAF1
SMAD3
IQCH
CRTC3
LITAF
RMI2
SOCS1
PRKCB
SULT1A1
SULT1A2
EIF3C
NUPR1
RABEP2
GDPD3
ITGAL
ADCY7
NOD2
PAPD5
CHTF8
ZFP90
PLCG2
CDH13
IRF8
EMC8
14.01
14.02
14.03
15.01
15.02
15.03
15.04
16.01
16.02
16.03
16.04
16.05
16.06
16.07
16.08
16.09
Chr.14
Chr.15
Chr.16
K
Mean FPKM
200
180
160
140
120
100
80
60
40
20
0
LGALS9
NOS2
CCL13
CCL2
CCL11
GSDMB
ORMDL3
IKZF3
STAT3
VPS25
PSME3
TRIM25
STAT5B
STAT5A
RPS6KB1
TUBD1
SLC39A11
CYTH1
USP36
PTPN2
SEC11C
SMAD7
CD226
ATP9B
17.01
17.02
17.03
17.04
17.05
17.06
17.07
17.08
18.01
18.02
18.03
18.04
18.05
Chr.17
Chr.18

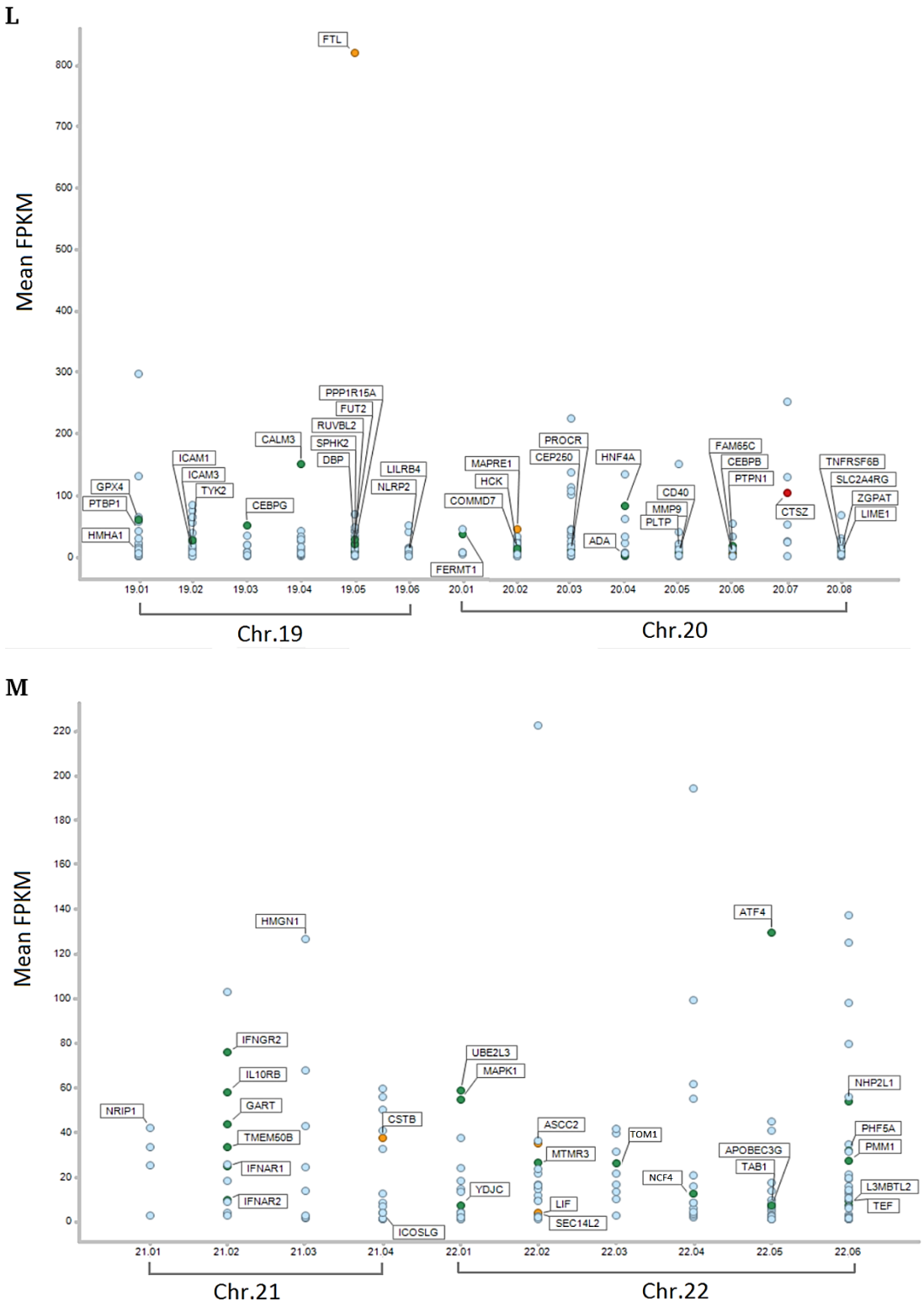


Figure 4.6 | Gene expression versus IBD susceptibility locus locations
Gene expression levels (FPKM) for all transcripts mapped to within 500Kb of the 224 known IBD susceptibility loci plotted per locus with genomic locations given in appendix 5 (A-M). Each dot represents a transcript with detectable expression in the intestinal RNAseq data. The spots are colour to indicate how they have been implicated as potentially important in IBD. Green spot = previously prioritised by other studies; yellow spot = is differentially expressed in IBD vs controls in the RNAseq data, red = both differentially expressed and prioritised; blue = all other non-prioritised transcripts.

Figure 4.6 shows observed gene expression at all 224 IBD loci with previously prioritised genes and DE genes highlighted. It was noted that, whilst some previously prioritised genes showed no expression above background levels (>1 FPKM) within the colon RNAseq data, others contained multiple highly expressed and previously prioritised genes e.g. locus 1.18 (**Figure 4.6**). However, it was also observed that 67 loci contained multiple expressed genes but only one of the previously prioritised or differentially expressed genes were expressed above background e.g. locus 5.01 (**Figure 4.6**). Furthermore, locus 6.03 was observed to contain 9 transcripts but none reached expression above background.

Knowledge about previously prioritised functional candidate genes for IBD was combined with the generated qualitative and differential expression data from the large intestine. A list of 554 genes, spanning 179 IBD susceptibility loci, which are potentially involved in IBD pathogenesis was established. Eight regions were identified where the previously proposed functional candidate gene showed altered expression in colonic samples from IBD patients and a further 64 regions were identified where only one prioritised gene shows detectable expression in these tissues.

4.5 Discussion

Recent progress in prioritising causal genes in IBD has been made through fine-mapping and expression quantitative trait (eQTL) analysis, with 16 of the known IBD susceptibility loci having been reduced to 1 causal variant at >95% probability [108,112]. A further 445 genes have been prioritised to be involved in IBD pathogenesis based on investigation of gene co-localisation, protein-protein interactions, functional connectivity of a gene within literature and eQTL analysis [107,150]. Here we aimed to increase the knowledge of the pathogenesis of CD and IBD by characterising the entire colonic transcriptome, using whole RNA sequencing, and investigate differences in gene expression at IBD susceptibility loci in biological relevant intestinal tissue from affected patients and controls.

The patient cohort consisted of 75 CD patients, 24 UC cases and 28 controls. The imbalance in patient numbers was caused by a shift in focus of this study from CD only to include UC and controls, as well as by the logistics of sample collection; a larger number of CD patients have colonoscopies compared to UC and controls. EdgeR has been designed to take into account sample imbalance, nevertheless it should be taken into consideration that we have a substantial sample imbalance and it might exhibit an effect on the differential expression analysis. It was decided to use un-inflamed large intestinal tissue over inflamed tissue. Inflamed tissue samples show stronger expression signals of immunological and/or pro-inflammatory IBD-associated genes, but it is hard to distinguish if these genes are upregulated due to the primary cause of disease or just secondary to the inflammation that results. Using un-inflamed intestinal tissue will slightly reduce strength of the signal and it will exclude any inflammatory signals but it will allow the separation between primary and secondary effects. Ideally, future studies will include both un-inflamed and inflamed tissue biopsies from IBD patients to address this limitation. Furthermore, the intestinal biopsies used to generate transcriptional data (see Material and Methods section 2.2.2) consist of a heterogeneous tissue including epithelial, stromal and various immune cell types. Expression signals

measured within heterogeneous tissues are confounded by relative proportions of the cell types involved, making it challenging to determine whether variability in gene expression stemmed from differences in phenotype or tissue composition. Various methods to resolve tissue heterogeneity have been proposed including *in silico* approaches to address tissue heterogeneity and purification of the cell populations through flow cytometry based cell sorting. Both methods have their limitations, which are discussed in more detail in Chapter 7. We attempted to address tissue heterogeneity of the intestinal biopsy samples by investigating a method for deconvolution of biopsy composition by utilising the gene expression data.

4.5.1 Gene expression within the colon

When initiating this study gene expression data was starting to emerge using microarray analysis but no whole RNA sequencing data from large intestinal tissue of healthy controls or IBD patients had been published. Since then, two studies investigating the influence of diet and lifestyle on the transcriptome of colonic tissue in healthy controls and colorectal cancer patients have been published [165,166]as well as an in-depth microarray study mapping the gene expression landscape including lncRNA and mRNA in colonic tissue of healthy controls and IBD cases [167]. The first aim of this study was to examine expression levels of genes and non-coding RNAs within uninflamed large intestinal tissue, specifically within the IBD susceptibility loci, of an n > 100 individuals. Although, the above mentioned studies investigated the transcriptome in colonic tissue they did not publish an overall number of genes found to be expressed within the colon or specifically within IBD susceptibility loci locations [165-167]. Within the transcriptomic data generated in this study, 56,360 transcripts were observed to align to the reference genome within the colonic samples, of which 32% exhibited expression above background (≥1 FPKM). Of these, 77% were coding genes and approximately 3% were lncRNAs. The remaining 20% included pseudogenes, processed transcripts and small RNAs. It was observed that 2,971 transcripts within 500Kb of 224 known IBD loci

exhibited expression above background. Approximately, 19% of genes previously prioritised to be involved in IBD pathogenesis were not expressed above background in the uninflamed intestinal tissue data, including cytokines IL3, IL4, IL5, IL12B, IL13, IL19, IL20, IL21, IL22, IL26, IL27 and IL31RA [107,150]. These cytokines had been prioritised using GRAIL connectivity network analysis. Their failure to reach expression levels above background in the generated uninflamed colonic tissue dataset could suggest their involvement in IBD is due to secondary effects or their effect on IBD will only be measurable within inflamed tissue. For this study it was hypothesised that sequencing and mapping of all transcripts expressed within known IBD loci, would provide further insight into overall gene expression in this disease relevant tissue and help provide more explicit functional evidence to determine which of the many previously prioritised genes are potentially causal.

4.5.2 Differential gene expression analysis

Beyond quantitative evaluation of transcription within colonic tissue at the IBD susceptibility loci, the aim of this study was to further the understanding of which genes contribute to IBD pathogenesis by comparing colonic gene expression levels between CD cases and controls, IBD and controls, UC and controls and UC cases vs CD cases.

4.5.2.1 CD *versus* control analysis

In our study 1,051 transcripts were shown to exhibit differential expression between CD cases and controls, 178 of which were located within known IBD loci. *DENND1B* (DENN Domain Containing 1B) exhibited reduced expression in CD cases vs controls. DENND1B has been shown to regulate T-cell receptor (TCR) internalization in Th2 cells, leading to an >3-fold increase of Th2 cytokines IL4, IL5 and IL13 following stimulation in DENN1B$^{-/-}$ mice [157]. Variation near the in *DENND1B* locus has been associated with various immune disorders including CD [168-170]. Using GRAIL analysis *DENND1B* was highlighted

as potentially important in CD [168], our observed decrease expression in CD cases supports to this hypothesis.

TNFRSF14 (Tumour Necrosis Factor Receptor Superfamily Member 14) expression was observed to be increased within colonic tissues of CD cases vs controls. TNFRSF14 (or HVEM) function in intestinal mucosa has been studied extensively because of its role in host defence and regulation of microbiota. TNFRSF14 is unique in its features; it can bind multiple ligands including non TNF super family members BTLA and CD160 and it can act as a receptor or ligand [171]. TNFRSF14 has been shown to induce activation of mucosal T cells when interacting with its ligand TNFSF14 (LIGHT) and regulate the epithelial innate immune responses to bacterial infection [172,173]. However, when interacting with BTLA, inhibition of T-cell activation was observed [171,174,175]. Upregulation of TNFRSF14 within CD patients could indicate an increased signalling through the LIGHT ligand resulting in an increased T cell response.

4.5.2.2 IBD *versus* control analysis

Further subgroup analysis of colonic expression in IBD cases vs control identified 526 transcripts to be differentially expressed, with 80 genes located in known IBD loci. The gene with the highest significant difference both within the CD and IBD vs control analysis was *GLS* (Glutaminase) which showed reduced expression within CD and IBD cases vs controls. Glutaminase is an enzyme involved in the hydrolysis of glutamine into glutamate and ammonia. The role of glutamine in the body and specifically in the intestine has been researched extensively. It has been shown that glutamine promotes immune cell functions including T cell proliferation, B cell differentiation, phagocytosis, antigen presentation and cytokine production [176-180]. Furthermore, reduced glutamine can lead to reduced gut mucosal integrity and increased gut permeability to allergens and pathogens [181]. A mouse study, showed increased bacterial translocations and intestinal permeability following introduction of radioactive labelled *E. coli* or diethylenetriamine pentaacetate (DTPA) into mice with or without glutamine in their diet [181]. Interestingly, it has been shown that the protective effect of glutamine does not require metabolism of

glutamine into glutamate [182]. The role of glutaminase is in the intestine is less defined, although it has been observed that both intestinal glutamine levels and glutaminase activity are reduced in CD patients [183]. The observed reduction of glutaminase levels in IBD patients collaborates our finding.

VIL1 (Villin1) was identified to exhibit reduced expression within colonic mucosa of IBD cases vs controls. VIL1 plays a role in intestinal cell morphology and cell migration [160]. In addition, increased expression of VIL1 has been suggested to protect against apoptosis of intestinal epithelium cells, with VIL1 knock-down mice with DSS-induced colitis showing increased severity of colitis and increased epithelial cell death [161,162]. Apoptosis is a tightly regulated process with a fine balance between necessary turn-over of epithelial cells and excessive levels of cell death leading to compromised epithelial barrier function [184]. The, in our study, observed reduced expression of *VIL1* could contribute to increased epithelial apoptosis and compromised barrier function in IBD patients.

4.5.2.3 UC *versus* control and UC *versus* CD analyses

When assessing the UC vs control DE results, q-values of a large subset of genes were observed to plateau out at approximately q = 0.2 (see Figure 4.3). Considering the DE analysis was performed on the smallest group of individuals (n=28 UC and n=24 controls), the plateau formation was contributed to reduced power in this subset. *TPSAB1* (Tryptase Alpha/Beta 1), showing reduced expression within UC cases, was the only gene to reach significance, although it is not located within one of the known IBD susceptibility loci. To overcome the lack of power and attempt to identify UC specific genes, an UC vs CD DE analysis was performed. The identified 134 genes, located within 500 kb of a known IBD locus, were thought to most likely represent unique differences between UC and CD phenotypes. Of these, 56 were previously found in the CD vs control analysis, the remaining 78 genes were investigated in UC pathogenesis. Top hits included, *TMEM259* (Transmembrane protein 259) and *GAL3ST2* (Galactose-3-O-Sulfotransferase 2), both exhibited lower expression in UC vs CD. TMEM259 may play a role in clearance of misfolded

proteins in the endoplasmic reticulum (ER) and has been suggested to promote survival of motor neurons. Further functional knowledge about TMEM259 is limited, with no obvious link to IBD or the immune responses. GAL3ST2 is known to be present in intestinal mucosa where it catalyses the sulfonication of mucins e.g. synthesises sulfomucins [185]. Sulfomucins have been implicated in protection of the intestinal mucosa and are suspected to enhance the mucus viscosity and resistance to bacterial degradation and microbe adhesion [185,186]. Studies have shown that a significant loss of sulfomucins can be observed in the mucosal lining of UC patients [187,188]. It is possible this loss of sulfomucins could be in some part due to the observed reduced colonic expression of *GAL3ST2*. Further investigation into this pathway is warranted in our cohort. The observed reduced expression of *GAL3ST2* within our UC patient could explain the reduced synthesis of sulfomucins.

5. Pathway analysis of genes differentially expressed in IBD

In the previous chapter, high quality colonic gene expression data from heathy controls and IBD patients (UC and CD), were subjected to differential expression (DE) analysis in order to identify genes that may play a role in the pathogenesis of IBD. Various sub-analyses were performed to identify transcripts contributing to IBD overall and CD or UC specific DE transcripts: IBD vs controls, CD vs controls, UC vs controls and UC vs CD, identifying 526, 1051, 1 and 696 DE transcripts, respectively. The UC vs control DE analyses was underpowered, resulting in the identification of only 1 DE gene. Each of the significant ($q \leq 0.05$) DE gene lists, with the exception of the UC vs controls DE genes, were advanced into pathway analysis to devise their potential influence on biological processes. To fully utilize the generated data two different pathway analyses were employed: gene set enrichment analysis (GSEA) and Ingenuity pathway analysis (IPA).

5.1 Pathway analysis tools

Pathway analysis combines statistical enrichment methods and prior knowledge of gene function to identify biological processes and molecular pathways affected by differentially expressed genes. Databases containing information on gene, protein, metabolite or compound interactions curated from the scientific literature are used to aid pathway identification. Here two independent methods were employed to investigate potential underlying biological pathways implicated in IBD pathogenesis through the DE analysis: Gene Set Enrichment Analysis (GSEA) [148], developed by the Broad Institute, and Ingenuity Pathway analysis (IPA), developed by Qiagen.

5.2 Gene Set Enrichment analysis (GSEA)

GSEA extracts biological insight from differentially expressed genes by comparing them to gene sets within the Molecular Signature Database (MSigDB) [148]. All gene sets are based on experimental outcomes and often

report common biological function, regulation or chromosome location. MSigDB contains, to date, 13,311 gene sets divided into 8 major collections. Out of the 8 available collection 3 were investigated:

- C2 containing 4,726 curated gene sets collected from various public online databases such as Biocarta (an interactive web-based resource for life sciences), KEGG (Kyoto Encyclopaedia of Genes and Genomes), REACTOME (a curated pathway database), Signalling Gateway and Pathway Interaction Database.
- C5 containing 1,451 gene ontology (GO) gene sets consisting of co-regulated genes.
- C7 containing 4,872 Immunological signature gene sets representing cell states and perturbations within the immune system [189].

GSEA pre-ranked analysis allowed the investigated genes to be ranked based on q-value and fold change for each gene (see Materials and Methods section 2.2.9.4). Genes with a low q-value together with a strong positive or negative fold change were ranked at the top or bottom of the ranked list, respectively. Enrichment for genes ranked towards either the top or bottom of the pre-ranked list was then assessed against the presence of these genes within gene sets recorded in the MSigDB. An enrichment score (ES) and false discovery rate (FDR) was calculated for each MSigDB gene set. Pathway analysis was performed on three of the DE gene sets: IBD vs controls, CD vs controls and UC vs CD. The UC vs control DE analysis was underpowered and therefore not taken forward.

5.2.1 IBD *versus* control GSEA

All 15,379 genes investigated for differential expression between IBD and controls were pre-ranked and run against 11,052 MSigDB gene sets (4,726 curated, 1,454 GO and 4,872 immunological gene sets) [148]. Overall, 3,429 gene sets showing enrichment amongst IBD differentially expressed genes were identified at a FDR of $\leq$ 5%. Of which, 2,329 (68%) were amongst the immunological gene sets in MSigDB. A two sided probability test indicated that

the enrichment in immunological signatures was significantly greater than would be expected by chance ($p < 2.2x10^{-16}$). From the 3,429 significantly enriched gene sets, 78 were positively enriched in IBD (containing genes upregulated in IBD) and 3,351 showed negative enrichment in IBD (containing genes downregulated in IBD).

Within the 3,351 negatively correlated gene sets, most notably, enrichment was observed within gene sets related to glucocorticoid therapy and glutamine deprivation (**Figure 5.1**).

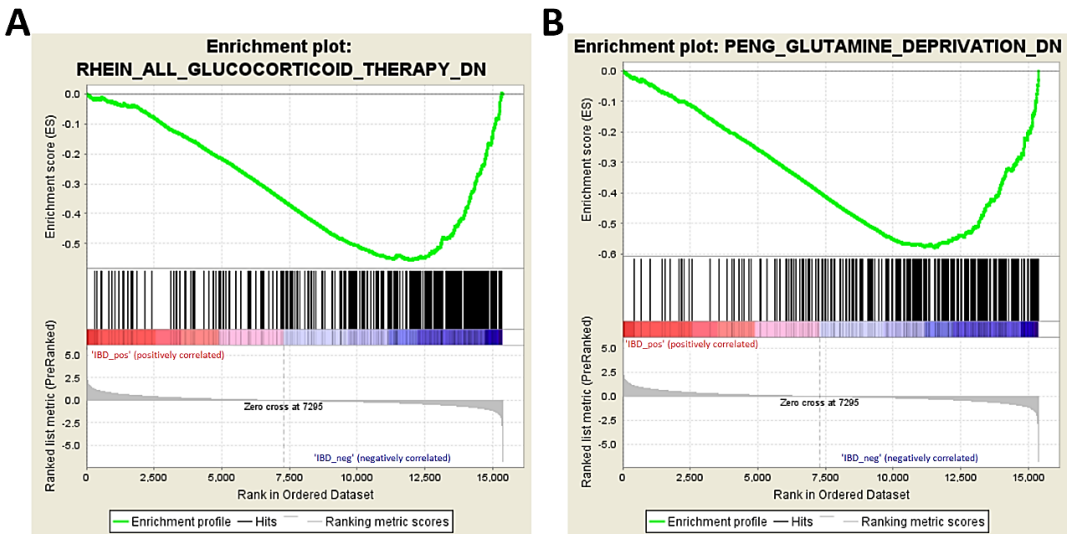


Figure 5.1 | Gene set enrichment plots
Gene set enrichment plots showing negative enrichment in IBD vs controls. The top half of the graph shows the enrichment score (green line) based on presence and weight of ranked genes (black lines in centre) within the pathway. Red indicating top of ranked list (positive correlation with IBD) and blue indicating negative correlation with IBD. The bottom half of the graph shows ranked list metric scores indicating weight of ranking versus the location of each gene (black lines in centre) within the ranked gene list. (The enrichment plot was generated by GSEA software [148], http://www.broad.mit.edu/gsea/).

Glucocorticoids (GCs) are a class of steroid hormones which reduce inflammation and are often used to treat autoimmune diseases including IBD. Approximately 340 genes downregulated in IBD were found in a gene set identified following glucocorticoid therapy for leukaemia [190]. Enrichment in this pathway suggests that gene expression changes in response to GCs may

overlap in leukemic precursor B-cells and intestinal tissue. In addition, negative enrichment within a set of 321 genes downregulated following glutamine deprivation suggested that glutamine metabolism might be perturbed in IBD patients. Glutamine has been proposed to have an essential function in the gut, specifically under stress. It has been shown to decrease gut permeability, prevent bacterial translocation, promote gut integrity and optimise nitrogen balance [191,192]. Furthermore, Reactome pathways including regulation of apoptosis, Wnt signalling, antigen cross presentation and adaptive immune system showed negative enrichment.

Gene sets positively correlated with IBD included, most notably, Interferon-alpha (IFNα) activated PBMCs, NCAM1 (Neural Cell Adhesion Molecule 1) interactions and signalling by Notch 4 (**Figure 5.2A-C**). The IFNα-activated PBMCs gene set showed the biggest overlap with the pre-ranked gene list at 86 genes (**Figure 5.2A**). The IFNα-activated PBMCs gene set was one of various gene sets comparing naïve vs activated immune cell state, including $CD4^{pos}$ T cells, $CD8^{pos}$ T cells, monocytes, macrophages, B cells and NK cells, suggesting an increased presence of activated immune cells in the IBD intestinal tissue vs controls. The NCAM1 interactions gene set showed enrichment for 27 genes more highly expressed in IBD (**Figure 5.2B**). NCAM1 interactions include cell adhesions, proliferation, differentiation, migration and cell survival. Enrichment in signalling by two Notch proteins was observed; Notch 3 and Notch 4 (**Figure 5.2C-D**), although only Notch 4 reached FDR $\leq$ 5%. Enrichment in signalling by Notch 3 and Notch 4 were driven by the presence of 12 DE genes, identical in both Notch 3 and Notch 4. Notch 3 and 4 signalling is reported to be activated by delta-like and jagged ligands (DLL/JAG) resulting in transcriptional changes by the notch intracellular domain. The Notch protein family is known to regulate a broad spectrum of cell fate decisions. A further subset of positively correlated gene sets were involved in in cell polarisation and channel activity including voltage gated channel activity, cation and calcium channel activity.

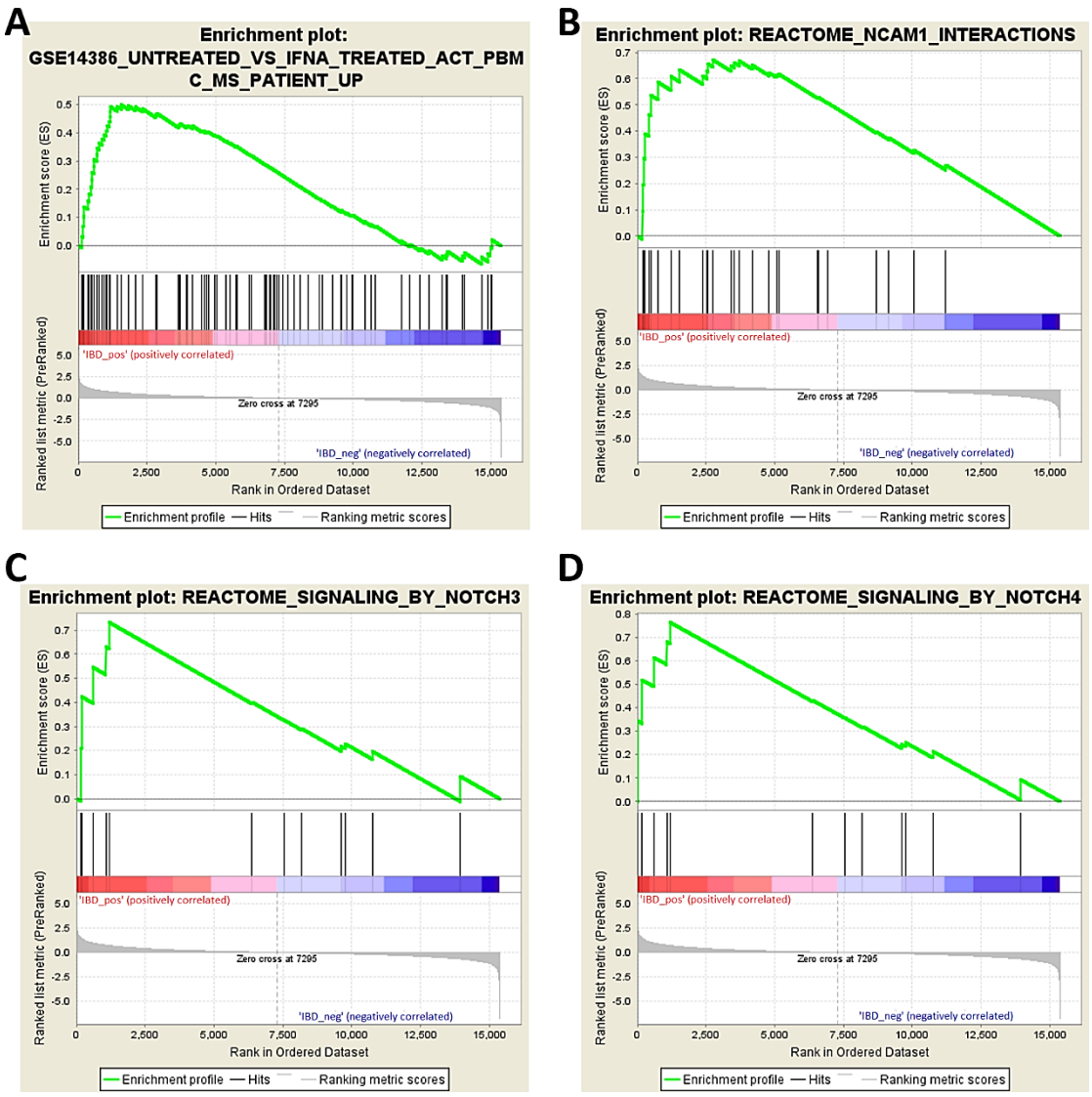


Figure 5.2 | Gene set enrichment plots
Gene set enrichment plots showing negative enrichment in IBD vs controls. The top half of the graph shows the enrichment score (green line) based on presence and weight of ranked genes (black lines in centre) within the pathway. Red indicating top of ranked list (positive correlation with IBD) and blue indicating negative correlation with IBD. The bottom half of the graph shows ranked list metric scores indicating weight of ranking versus the location of each gene (black lines in centre) within the ranked gene list. (The enrichment plot was generated by GSEA software [148], http://www.broad.mit.edu/gsea/).

5.2.2 CD *versus* control GSEA

All genes investigated for differential expression between CD and controls were pre-ranked and run against 11,052 MSigDB gene sets (4,726 curated, 1,454 GO and 4,872 immunological gene sets) [148]. In total, 1,823 gene sets showed enrichment at a FDR of ≤ 5% with a significant overrepresentation ($p < 2.2e^{-}$

[16]) of gene sets belonging to the immunological gene set in MSigDB (74%). Out of the 1,823 significantly enriched gene sets, 121 were positively enriched in CD and 1,702 showed negative enrichment in CD. Interestingly, the CD vs control analysis compared to the IBD vs control analysis, identified double the number of differentially expressed genes (1,051 vs 526) and only half the number of pathway enrichment signals (1,823 vs 3,429). Of the 1,823 enriched gene sets identified 1,718 were observed to also be enriched in IBD vs controls, resulting in 105 newly identified enriched gene sets, 46 negatively and 59 positively.

The 46 significant negatively enriched gene sets included 8 gene sets involved in macrophage activation, 2 metabolism pathways, steroid hormone biosynthesis pathways and a pathway involved in xenobiotics metabolism (**Figure 5.3A**). Interestingly, negative enrichment was observed within a gene set identifying changes in gene expression within keratinocytes following exposure to UV (**Figure 5.3B**).

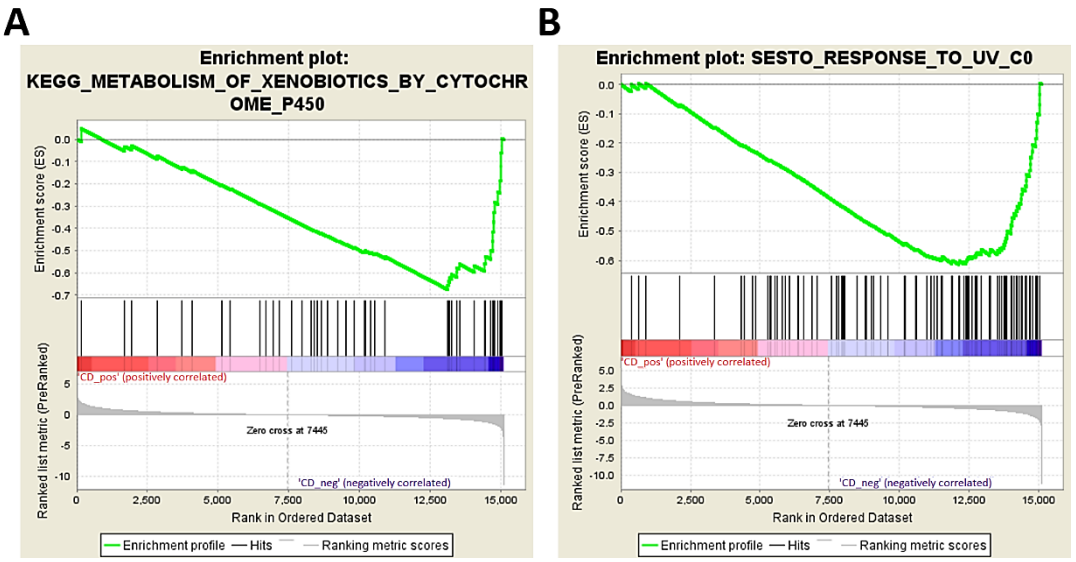


Figure 5.3 | Gene set enrichment plots
Gene set enrichment plots showing negative enrichment in CD vs controls. The top half of the graph shows the enrichment score (green line) based on presence and weight of ranked genes (black lines in centre) within the pathway. Red indicating top of ranked list (positive correlation with CD) and blue indicating negative correlation with CD. The bottom half of the graph shows ranked list metric scores indicating weight of ranking versus the location of each gene (black lines in centre) within the ranked gene list. (The enrichment plot was generated by GSEA software [148], http://www.broad.mit.edu/gsea/).

Sesto *et al.* identified 9 gene sets showing changes in expression in response to UVB, including upregulated genes involved in UV-specific inflammatory and stress responses as well as downregulated genes involved in metabolism and adhesion processes [193]. Upon further investigation enrichment ($q \leq 0.05$) in 5 out of 9 of these identified gene sets was observed. Four UVB response gene sets were identified in both IBD and CD vs control analysis and one was identified in only CD vs controls. Thiopurines, most commonly Azathioprine, have been associated with increased risk to non-melanoma skin cancer, enrichment of these gene sets might be an effect caused by the thiopurine drugs taken by a subset of patients.

Within the 59 identified positively enriched gene sets, enrichment within the KEGG Notch signalling pathway was observed (**Figure 5.4A**). Signalling by Notch 3 and Notch 4 was observed to be enriched within the IBD vs control analysis (12 DE genes); their enrichment has been replicated in the

CD vs control analysis. Furthermore, the overall Notch signalling pathway was shown to be enriched, with 43 DE genes associated with this pathway. Notch proteins have been reported to regulate a broad spectrum of cell fate decisions by functioning as a receptor for transmembrane ligands to Jagged and Delta-like proteins. Furthermore, enrichment in a gene set altered in PBMCs infected by *Escherichia coli* (*E. coli*) versus healthy PBMCs was observed (**Figure 5.4B**).

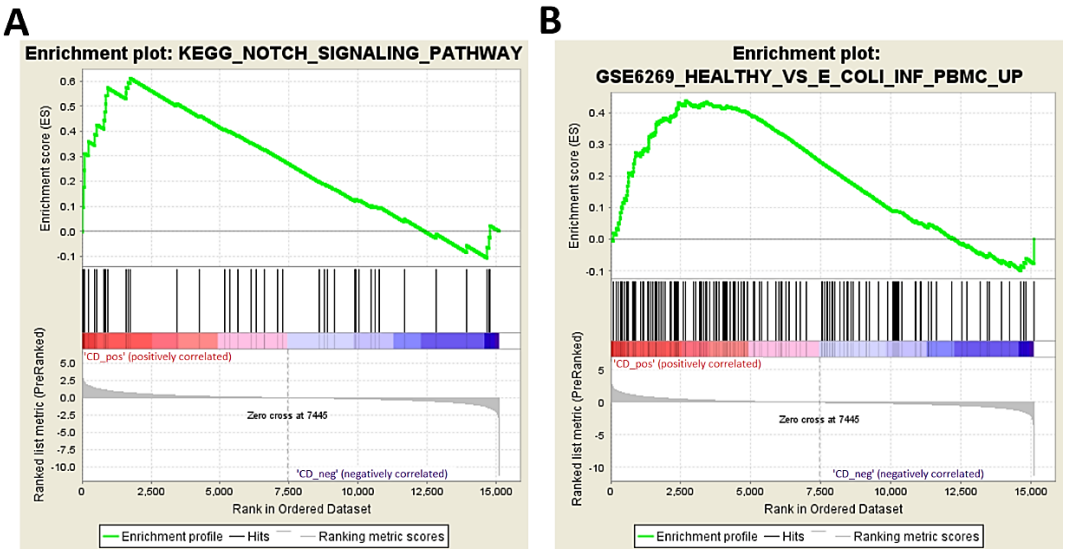


Figure 5.4 | Gene set enrichment plots
Gene set enrichment plots showing negative enrichment in CD vs controls. The top half of the graph shows the enrichment score (green line) based on presence and weight of ranked genes (black lines in centre) within the pathway. Red indicating top of ranked list (positive correlation with CD) and blue indicating negative correlation with CD. The bottom half of the graph shows ranked list metric scores indicating weight of ranking versus the location of each gene (black lines in centre) within the ranked gene list. (The enrichment plot was generated by GSEA software [148], http://www.broad.mit.edu/gsea/).

5.2.3 UC *versus* CD GSEA

The UC vs CD analysis was performed to identify pathways specific to either disease and not general to IBD. All genes investigated for differential expression between UC and CD were pre-ranked and run against 11,052 MSigDB gene sets (4,726 curated, 1,454 GO and 4,872 immunological gene sets) [148]. In total, 595 gene sets showed enrichment at a FDR of ≤ 5%. with a significant overrepresentation ($p < 2.2x10^{-16}$) of gene sets belonging to the immunological gene set in MSigDB (82%). Out of the 595 significantly enriched gene sets, 285

were enriched in UC (positive enrichment) and 310 showed enrichment in CD (negative enrichment). Notably, a more even spread between negative and positive enrichment within the UC vs CD analysis compared to either IBD or CD vs controls was observed.

An overlap of 223 enriched gene sets between all 3 analyses i.e. IBD and CD vs control and UC vs CD was detected (**Figure 5.5**). Of these, 217 showed opposite directions of enrichment in the UC vs CD analysis compared to the IBD and CD vs control analysis. The 6 gene sets showing negative enrichment in all three analyses were all observed within immunological cell types; regulatory T cells (Tregs), $CD4^{pos}$ T helper cells, $CD8^{pos}$ cytotoxic T cells and B cells, suggesting they might represent secondary inflammatory responses.

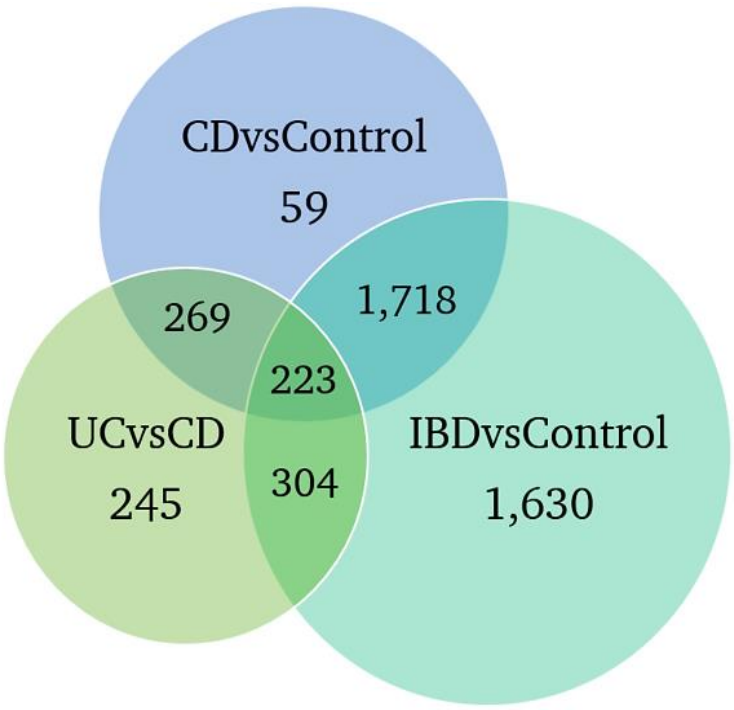


Figure 5.5 | Overlap GSEA analyses results
Venn diagram showing the overlap between the GSEA significant enrichment results of the IBD versus control (n=1,630 unique), CD vs control (n=59 unique) and UC vs CD (n=245 unique) analyses. With 223 enrichment gene sets showing overlap in all three results.

The UC vs CD analysis showed 245 gene sets to be uniquely enriched, with the most significant UC enriched gene sets observed to be involved in pathway activation processes within various cell types, providing insight into gene expression changes following immune response. For example, LPS (lipopolysaccharide) vs control IgG treated monocytes (**Figure 5.6A**) or IFNα- vs IFNγ-treated endothelial cells (**Figure 5.6B**) both showed positive enrichment. LPS, the outer cell membrane of gram-negative bacteria, is known to activate immune and inflammatory responses through TLR4 (Toll-like receptor 4) signalling [194].

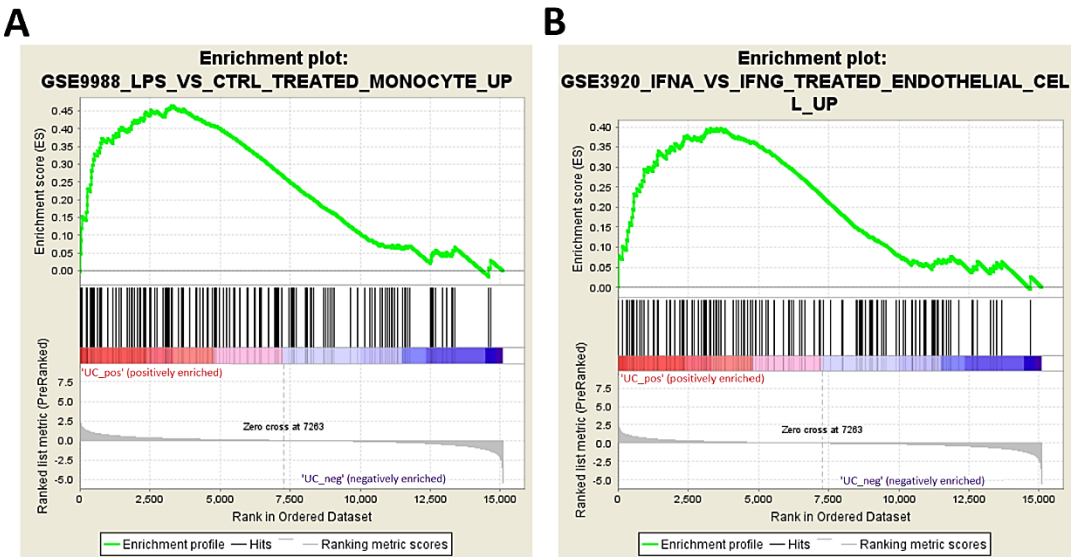


Figure 5.6 | Gene set enrichment plot
Gene set enrichment plots showing negative enrichment in UC vs CD. The top half of the graph shows the enrichment score (green line) based on presence and weight of ranked genes (black lines in centre) within the pathway. Red indicating top of ranked list (positive correlation with UC) and blue indicating negative correlation with UC. The bottom half of the graph shows ranked list metric scores indicating weight of ranking versus the location of each gene (black lines in centre) within the ranked gene list. (The enrichment plot was generated by GSEA software [148], http://www.broad.mit.edu/gsea/).

Negatively enriched pathways – showing increased expression within CD – exhibited a large overlap with gene sets enriched within the CD vs control analyses (65 out of 77). Pathways negatively enriched and unique to the UC vs CD analysis were 85% immune response related, including apoptosis and interferon activation pathways.

5.3 Ingenuity Pathway Analysis

Ingenuity pathway analysis (IPA®, QIAGEN Redwood City) utilises the Ingenuity knowledge base to provide insight into molecular and chemical interactions as well as cellular phenotypes and disease processes within a dataset. The Ingenuity knowledge base is built upon a wide range of published information including textbooks, reviews, biomedical literature and a variety of public databases. All this information is structured into a framework organising and describing biological evidence including contextual information; species specific info, cell type/tissue context, direction of change and experimental methods. IPAs strength lies in the quality control and structuring of information within the database as well as the fact that the knowledge base is updated weekly.

IPA input files contained gene names, fold change values, q-values and FPKM (Fragments Per Kilobase of exon per Million fragments mapped) expression values for all genes identified as differentially expressed. Subsequently, Ingenuity generates a p-value and ratio score for each pathway assessing the level of pathway perturbation. The knowledge base tissue specific information allowed us to test pathways involved specifically in the large intestine.

5.3.1 IBD *versus* control IPA

Differential expression analysis identified 526 transcripts that were differentially expressed between IBD cases and controls (see Chapter 4.2.1). IPA identified 17 pathways to be significantly more perturbed than by chance ($p \leq 0.05$) due to the observed differential expression within these genes (**Figure 5.7**). Pathways included Granzyme A signalling, Notch signalling, communication between innate and adaptive immune cells, Altered T and B cell signalling and Crosstalk between dendritic cells and natural killer cells (**Figure 5.7**).

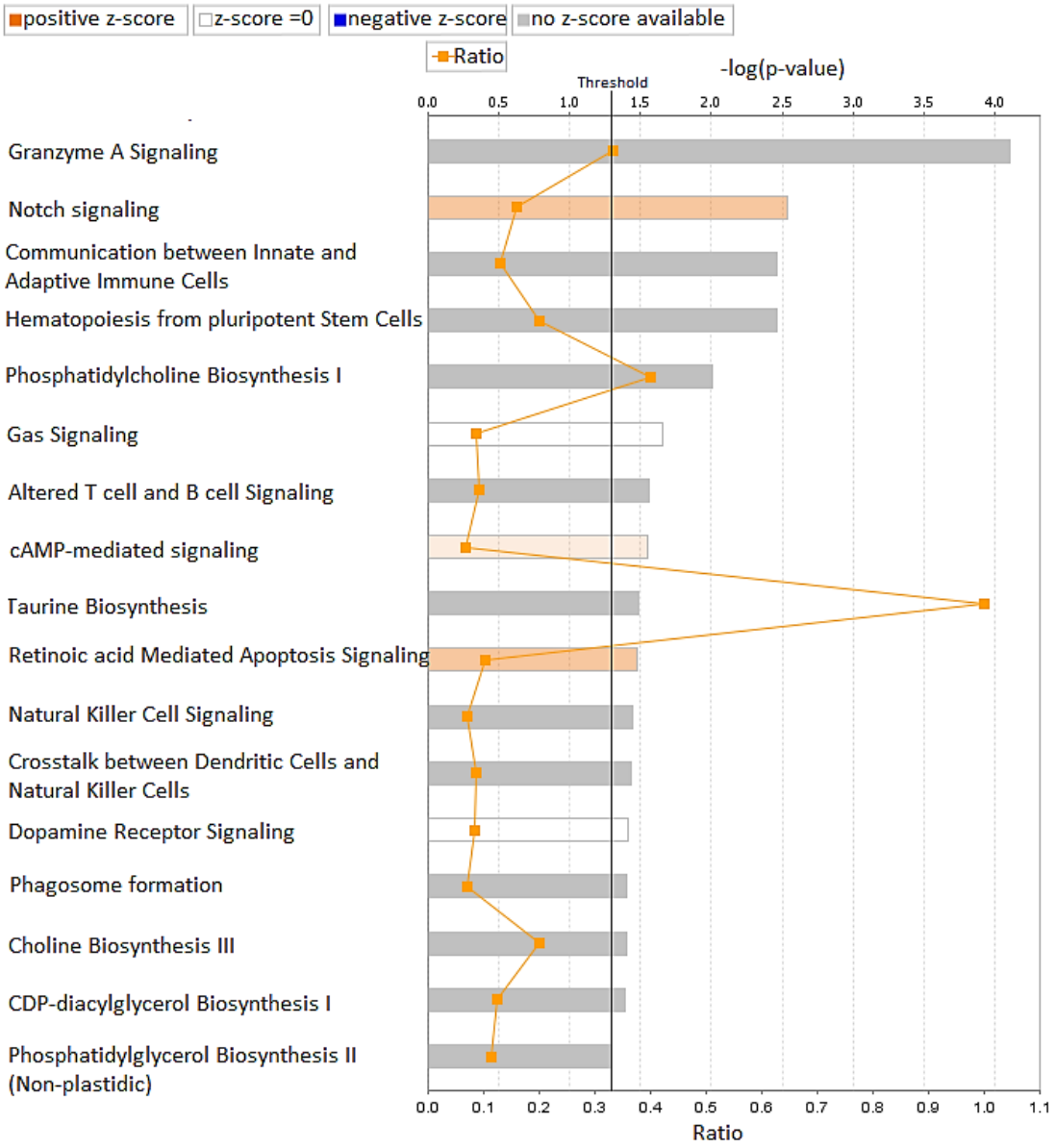


Figure 5.7 | IPA pathway analysis on colonic genes differentially expressed between IBD cases and controls
Pathways that are significantly more perturbed than by chance based on differentially expressed (DE) genes between IBD cases and controls. Bars represent the –log p value for the statistical test of number of genes DE in each pathway being more than expected by chance. Threshold indicated significance at p = 0.05. Bar shading indicates the direction of effect of DE genes on pathway activity with downregulated (blue) through to upregulated (orange), with grey indicating no information on direction of effect. Orange squares connected by the orange line indicate the ratio of the fraction of genes within a pathway which are significantly differentially expressed (q<0.05). (The pathway analysis was generated through the use of QIAGEN's Ingenuity Pathway Analysis (IPA®, QIAGEN Redwood City, www.qiagen.com/ingenuity).

The most significant pathway observed was Granzyme A signalling ($p = 7.8x10^{-5}$) with 33% of genes (5 out of 15) in the pathway differentially expressed in IBD (**Figure 5.7**). The observed effect is due to five genes: four H1 histone family genes and *HMGB2* (High Mobility Group Box 2) gene which encodes a DNA-binding protein (**Figure 5.8**). Although the direction of effect on Granzyme A signalling is not known (no z-score) (**Figure 5.7**), all 5 genes showed a negative fold change in expression in IBD biopsies compared to controls (**Figure 5.8**). Granzyme A is known to aid cytotoxic T lymphocytes and natural killer (NK) cells in response to cell infection, cancer transformation and antigen presentation (**Figure 5.8**).

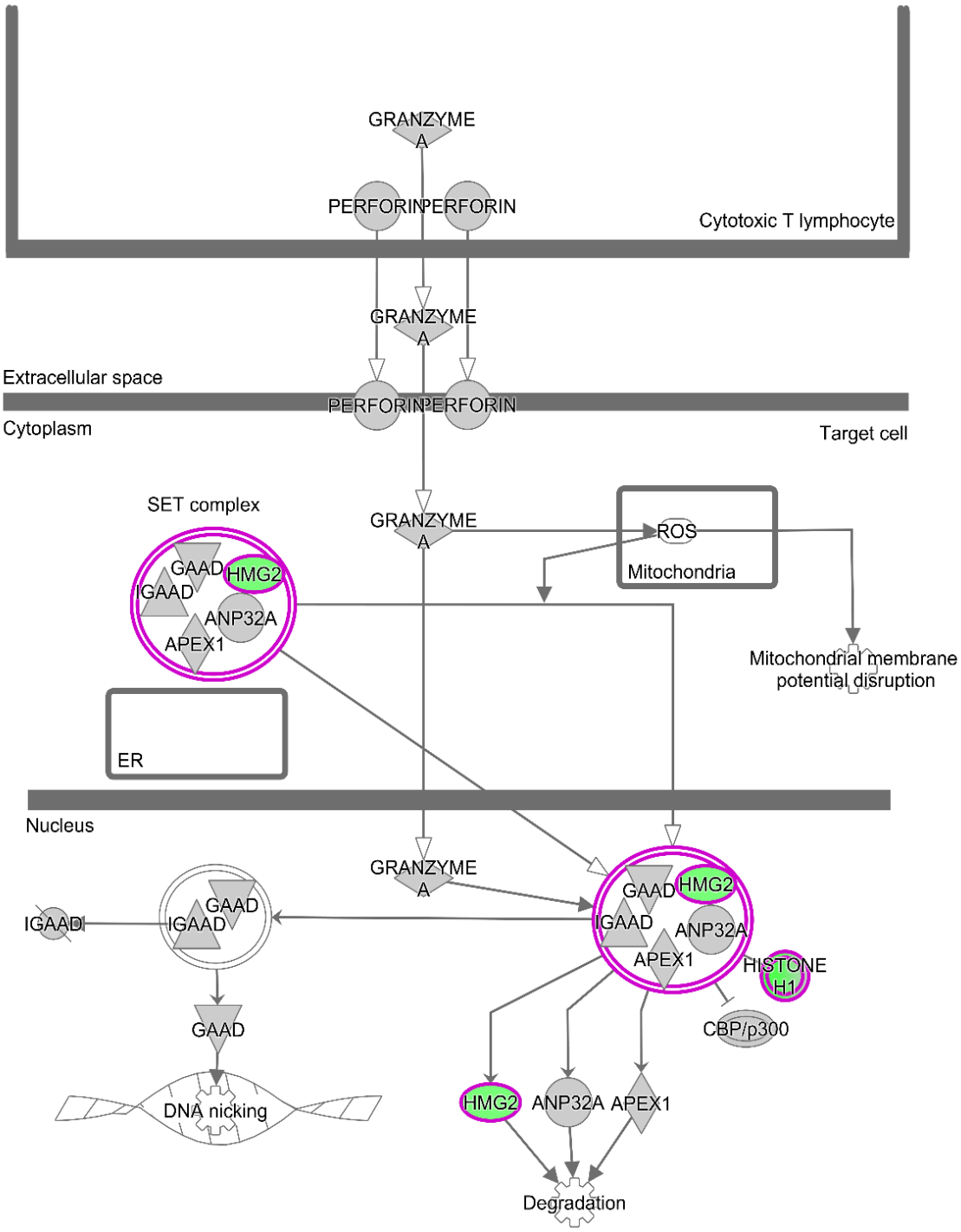


Figure 5.8 | Granzyme A signalling pathway
Granzyme A signalling pathway initiated within a cytotoxic T lymphocyte affecting a target cell, with double purple circles indicating protein complexes being affected by DE genes and green indicating downregulated DE genes. (The pathway analysis was generated through the use of QIAGEN's Ingenuity Pathway Analysis (IPA®, QIAGEN Redwood City, www.qiagen.com/ingenuity).

The Notch signalling pathway ($p = 2.9x10^{-3}$) was identified with 5 out of 31 genes within the pathway significantly differentially expressed (16%) (**Figure 5.9**). The 5 DE genes affecting the Notch signalling pathway were *DLL4* (Delta like canonical Notch ligand 4), *DTX2* (Deltex E3 ubiquitin ligase 2), *NOTCH3* (Notch 3), *NOTCH4* (Notch 4) and *NUMBL* (NUMB like). The positive z-score (**Figure 5.7**) indicated upregulation of pathway activity with all 5 genes showing increased expressed within IBD. Notch signalling has been implicated in maintenance of gut homeostasis and induction of UC when disturbed (**Figure 5.9**) [195,196]. Although the ratio of DE genes observed within the Notch signalling pathway was lower than within the Granzyme A signalling (33% vs 16%), the DE genes within the Notch signalling pathway appeared to affect the majority of protein complexes within the pathway (**Figure 5.9**).

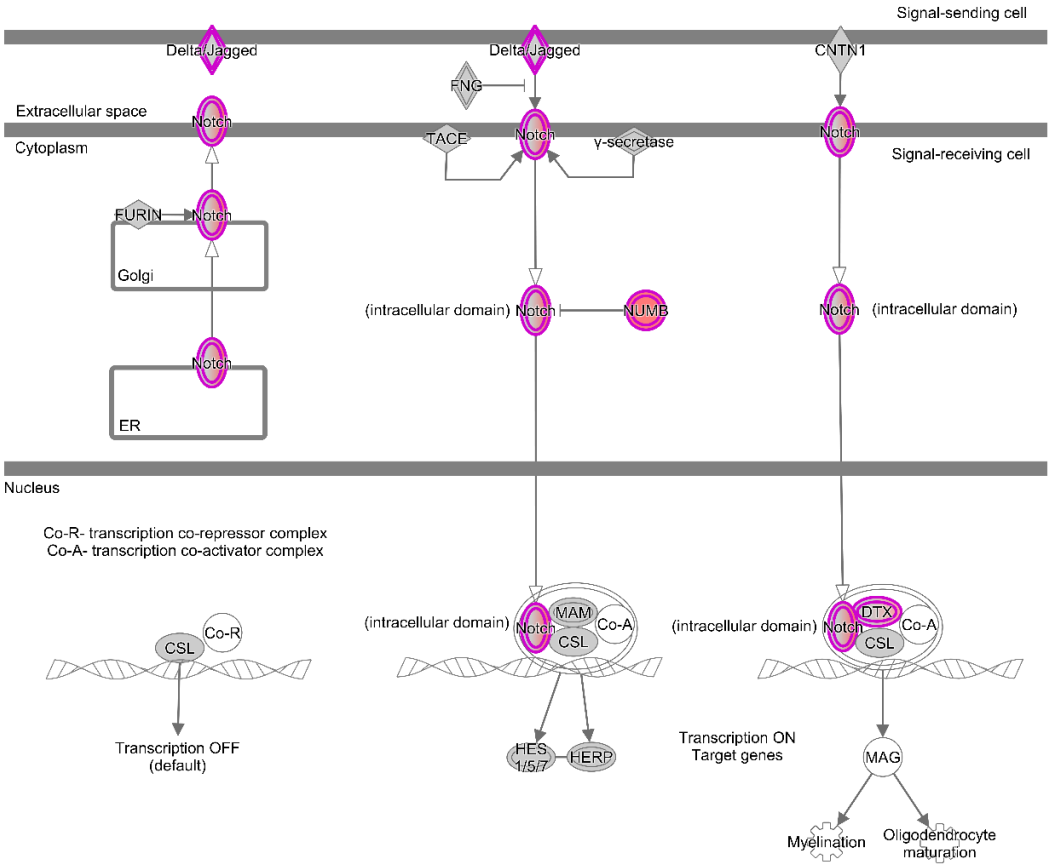


Figure 5.9 | Notch signalling pathway
Canonical pathway Notch signalling pathway showing both signalling and receiving cells. With purple double circles indicating protein complexes containing DE genes with the fill-in colour of the circle indicating direction of effect. NUMB indicated to be strongest upregulated (red). (The pathway analysis was generated through the use of QIAGEN's Ingenuity Pathway Analysis (IPA®, QIAGEN Redwood City, www.qiagen.com/ingenuity).

5.3.2 CD *versus* control IPA

CD and control DE analysis identified 1,051 transcripts that were differentially expressed (see Chapter 4.2.2). Within these genes IPA identified 28 pathways to be significantly more perturbed than by chance ($p < 0.05$), with Figure 5.10 showing the 20 most significant pathways.

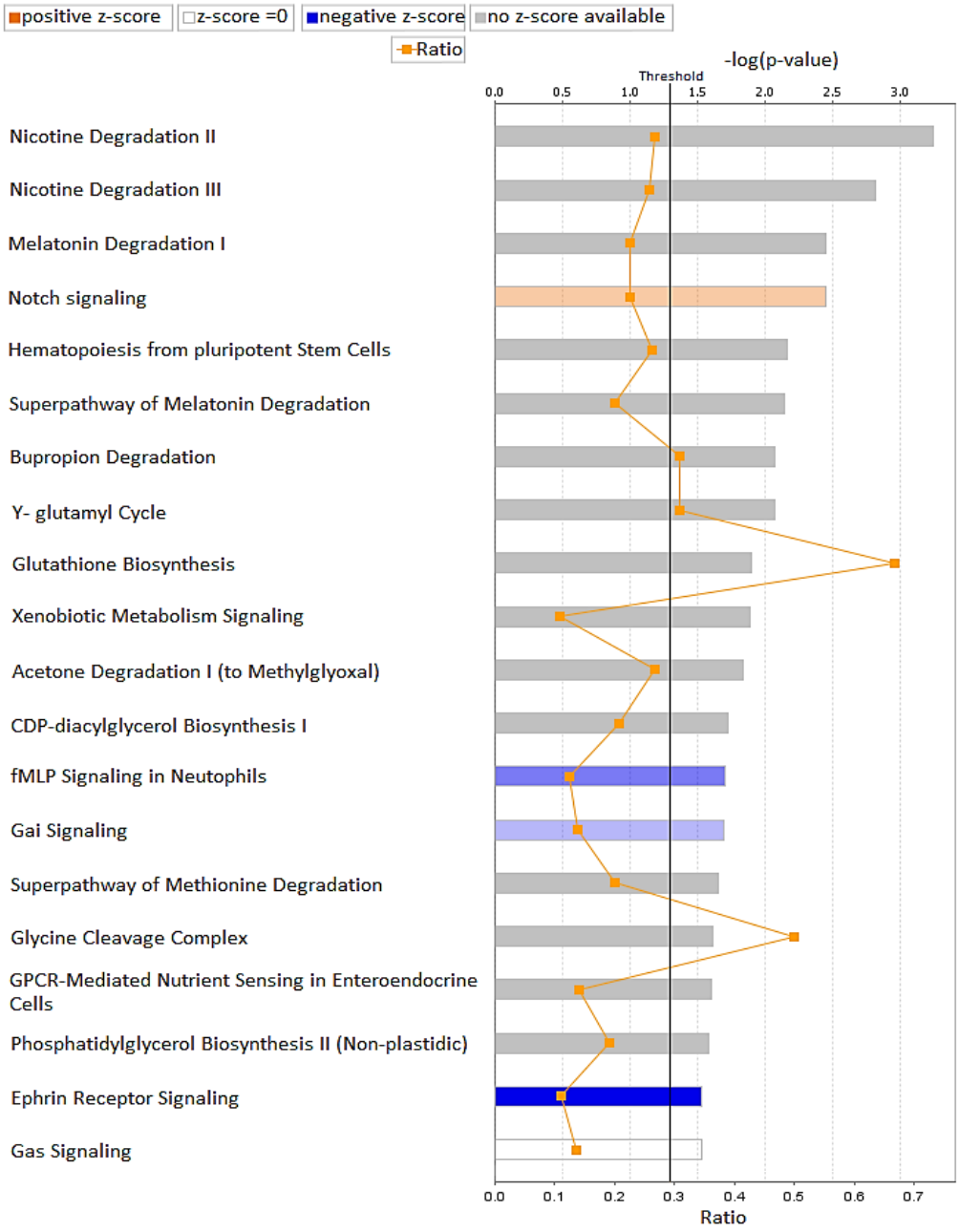


Figure 5.10 | IPA pathway analysis on colonic genes differentially expressed between CD cases and controls
Pathways that are significantly more perturbed than by chance based on differentially expressed (DE) genes between CD cases and controls. Bars represent the –log p value for the statistical test of number of genes DE in each pathway being more than expected by chance. Threshold indicated significance at $p = 0.05$. Bar shading indicates the direction of effect of DE genes on pathway activity with downregulated (blue) through to upregulated (orange), with grey indication no information on direction of effect. Orange squares connected by the orange line indicate the ratio of the fraction of genes within a pathway which are significantly differentially expressed ($q \leq 0.05$). (The pathway analysis was generated through the use of QIAGEN's Ingenuity Pathway Analysis (IPA®, QIAGEN Redwood City, www.qiagen.com/ingenuity).

The top two pathways suggested to be affected were Nicotine Degradation II ($p = 5.7x10^{-4}$) and Nicotine Degradation III ($p = 1.51x10^{-3}$). An overlap of 7 DE genes (**Table 5.1**) was observed between the two pathways. Of these, 4 were reported to be part of the Cytochrome P450 superfamily, known to be involved in catalysing drug metabolism [197]. The remaining 3 genes are part of the UDP glucuronosyltransferase family, known to be involved in transforming small lipids into excretable metabolites [198]. Although no z-score was recorded for either Nicotine degradation pathway, the 7 genes involved in both pathways all showed reduced expression within CD cases. Nicotine degradation II is reported to be involved in breaking down nicotine into various metabolites, one of which is cotinine (**Figure 5.11**). Nicotine degradation III has been observed to work downstream from nicotine degradation II and further breakdown cotinine into metabolites (**Figure 5.11**).

Table 5.1 | Gene functions of DE genes within Nicotine Degradation pathway

Gene Symbol	Family	Gene Function
CYP2B6 *CYP2C9* *CYP2C18* *CYP2C19*	Cytochrome P450 family 2	Catalysing drug metabolism
UGT1A4 *UGT1A6* *UGT1A7*	UDP glucuronosyltransferase family 1	Transformation of small lipophilic molecules into water-soluble, excretable metabolites

These 7 genes were observed to affect nearly every arm of both nicotine degradation pathways (**Figure 5.11**) suggesting they might exhibit a large effect on the degradation and metabolism of nicotine within CD cases. It was hypothesised that the observed effect might be due to a larger population of smokers within the CD cohort. The differential expression analysis was therefore repeated now including smoking status (current smoker, ex-smoker

or never smoked), as a covariate. The 7 genes driving the perturbation of the nicotine degradation pathway maintained their differential expression within the smoking-corrected CD vs control analysis (**Table 5.2**). IPA analysis likewise confirmed that both nicotine degradation pathways (II and III) remained perturbed, indicating that the observed effect cannot be attributed to smoking status, but to CD-specific gene expression patterns.

Table 5.2 | DE genes within Nicotine Degradation pathway

	Original analysis				Smoking corrected analysis		
Genes	**P-value**	**Q-value**	**Fold-change**		**P-value**	**Q-value**	**Fold-change**
CYP2B6	1.0E-01	4.4E-01	-0.50		1.1E-01	4.8E-01	-0.48
CYP2C9	8.4E-04	8.5E-02	-0.99		3.7E-03	1.5E-01	-0.85
CYP2C18	8.6E-04	8.5E-02	-0.75		1.7E-03	1.2E-01	-0.70
CYP2C19	4.6E-04	7.6E-02	-0.95		1.5E-03	1.1E-01	-0.85
INMT	4.4E-04	7.5E-02	0.38		3.8E-04	8.0E-02	0.36
UGT1A4	2.6E-01	6.1E-01	-0.20		1.7E-01	5.5E-01	-0.24
UGT1A6	9.1E-02	4.2E-01	-0.29		8.4E-02	4.4E-01	-0.29
UGT1A8	2.3E-02	2.6E-01	-0.35		4.4E-02	3.5E-01	-0.30

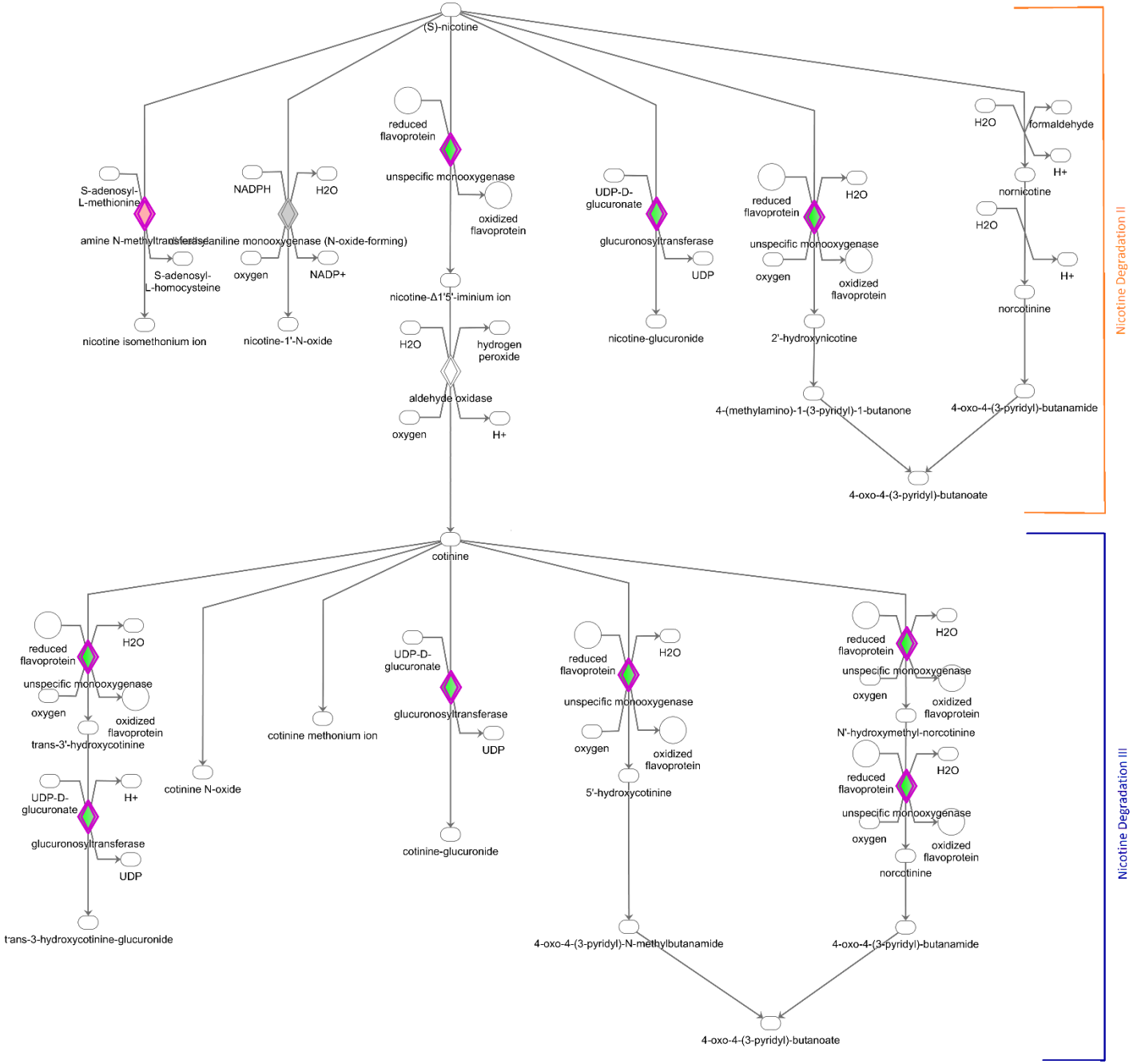


Figure 5.11 | Nicotine Degradation II and III Canonical pathways
Showing Nicotine Degradation II and III. Purple double circles indicating protein complexes containing DE genes with the fill-in colour of the circle indicating direction of effect: downregulation (green) to upregulation (red). (The pathway analysis was generated through the use of QIAGEN's Ingenuity Pathway Analysis (IPA®, QIAGEN Redwood City, www.qiagen.com/ingenuity).

The 7 genes from the nicotine degradation pathways (**Table 5.1**) were identified to be involved in an additional 3 significantly perturbed pathways; Melatonin Degradation I, Superpathway of Melatonin Degradation and Xenobiotic Metabolism. Moreover, the 4 Cytochrome P450 family genes also drove the observed significant enrichment within one of the other implicated pathways (Bupropion Degradation and Acetone degradation) and the 3 UDP glucuronosyltransferase family genes were identified within the Serotonin degradation pathway. The highest number of DE genes (23) was observed within the Xenobiotic Metabolism pathway, with the majority (16) indicating reduced expression levels within CD cases. These findings indicate that large portion of the observed pathways perturbed in CD tissues was due to a subset of genes involved in drug metabolism.

Ephrin Receptor signalling ($p = 2.9x10^{-2}$) exhibited a strong downregulated effect (negative z-score) due to 16 DE genes within the pathway (**Figure 5.10**). Ephrin receptor signalling has been shown to be involved with cell morphology, integrin-mediated adhesion and cell migration. Furthermore, pathways including Notch signalling and Gas signalling were identified in both IBD and CD vs control comparisons, with similar direction of effect and p-values.

5.3.3 UC *versus* CD IPA

To investigate IBD disease sub-type specific effects differential expression analysis between UC and CD was performed (see Chapter 4.2.4). The 696 transcripts that exhibited differential expression were investigated for their involvement and effect upon canonical pathways using IPA. In total, 21 pathways were observed to be significantly more perturbed than by chance ($p \leq 0.05$) (**Figure 5.12**).

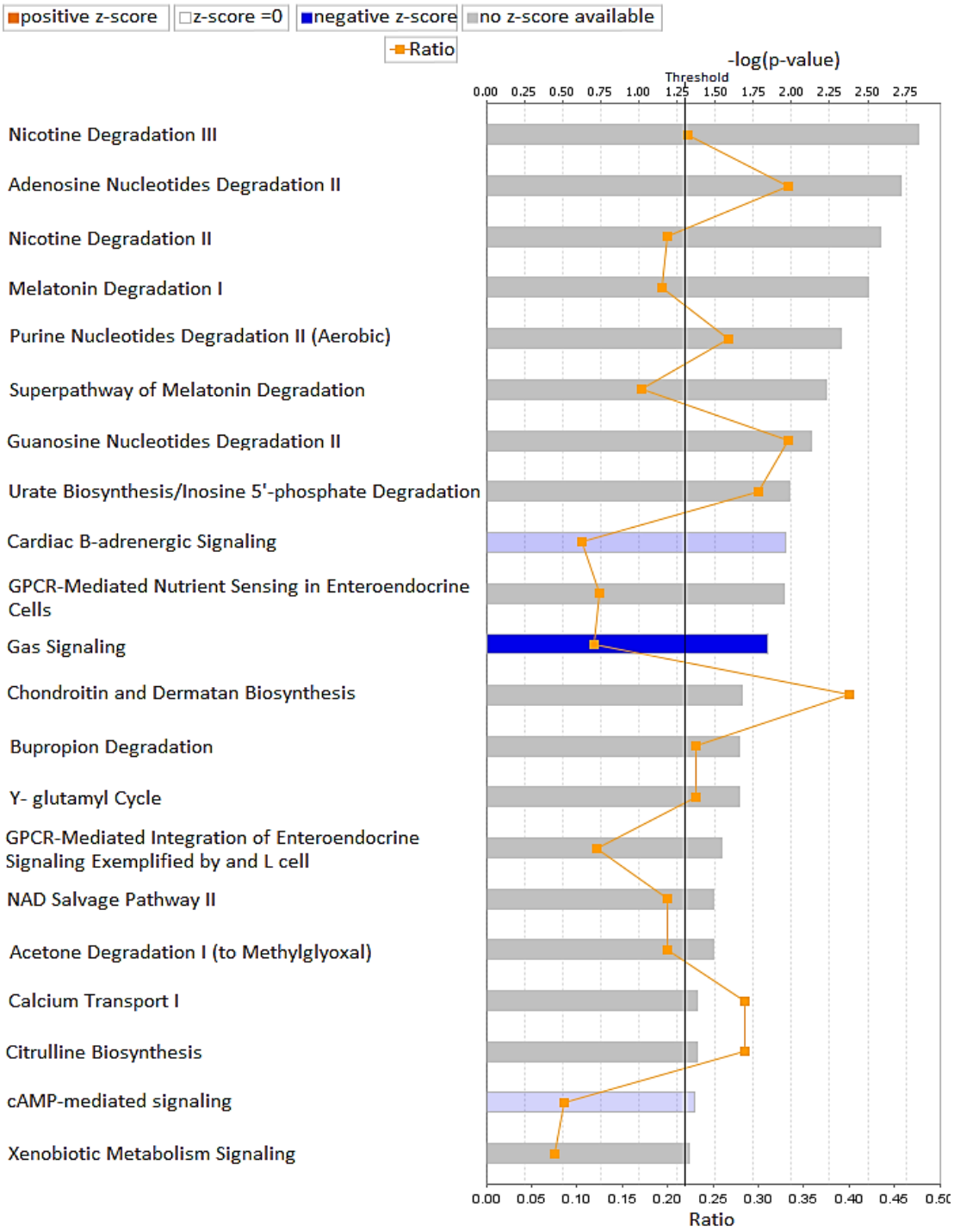


Figure 5.12 | IPA pathway analysis on colonic genes differentially expressed between UC and CD cases
Pathways that are significantly more perturbed than by chance based on differentially expressed (DE) genes between UC and CD cases. Bars represent the –log p value for the statistical test of number of genes DE in each pathway being more than expected by chance. Threshold indicated significance at p = 0.05. Bar shading indicates the direction of effect of DE genes on pathway activity with downregulated (blue) through to upregulated (orange), with grey indication no information on direction of effect. Orange squares connected by the orange line indicate the ratio of the fraction of genes within a pathway which are significantly differentially expressed ($q \leq 0.05$). (The pathway analysis was generated through the use of QIAGEN's Ingenuity Pathway Analysis (IPA®, QIAGEN Redwood City, www.qiagen.com/ingenuity).

Perturbed pathways identified in the CD vs controls and UC vs CD analyses showed an overlap of 10 affected pathways that were not implicated using DE genes identified in the IBD vs control analysis. Most notably, Nicotine Degradation II and III, Melatonin Degradation and GPCR-Medicated Nutrient Sensing in Enteroendocrine Cells (**Figure 5.13**).

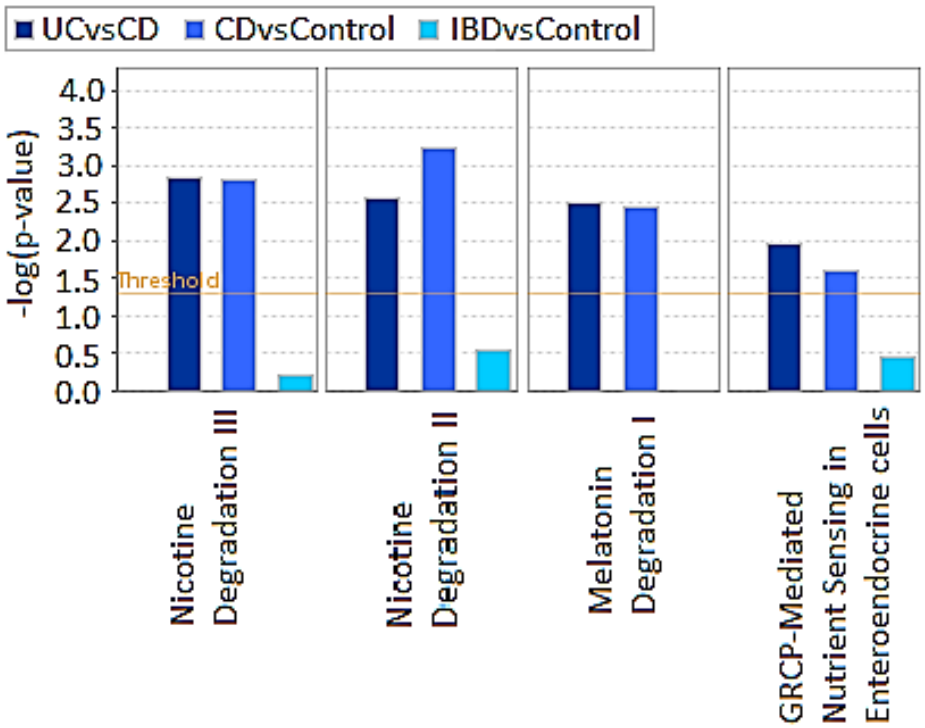


Figure 5.13 | Comparison subset IPA results
Bar charts plotting the -log(p-value) indicating levels of perturbation of a pathways based on differentially expressed genes versus a subset of pathways identified using Ingenuity Pathway analysis for UCvsCD, CDvsControl and IBDvsControl. Threshold (orange line) indicated significance at p = 0.05. (The pathway analysis was generated through the use of QIAGEN's Ingenuity Pathway Analysis (IPA®, QIAGEN Redwood City, www.qiagen.com/ingenuity).

The data summarised in Figure 5.13 suggest that the observed effects within these pathways may be specific to CD pathogenesis, although the consideration needs to be made that there was an imbalance in numbers for CD and UC cases (75 and 28, respectively). Similarly, it was speculated that the effect observed in pathways only affected by the UC vs CD DE gene list could be attributed to UC pathogenesis. Three such pathways were Adenosine Nucleotide Degradation II, Purine Nucleotide Degradation II and Guanosine Nucleotide Degradation III. Observed effects were all driven by decreased expression of

ACPP (Acid Phosphatase, Prostate), *ADAT3* (Adenosine Deaminase, TRNA specific 3) and *NT5C3B* (5'-Nucleotidase, Cytosolic IIIB), and increased expression of *NT5C2* (5'-Nucleotidase, Cytosolic II) in UC cases. Functions associated with these pathways include infectious diseases, cell signalling and cell death and survival. Two further pathways solely observed within the UC vs CD comparison were urate biosynthesis and NAD salvage pathway; when investigated it was observed that these were also driven by changes in expression of *ACPP*, *NT5C2* and *NT5C3B* genes. Gas signalling was the only pathway observed to be significantly affected in all three IPA analyses. Gas signalling was observed to be downregulated (negative z-score) within the UC vs CD comparison (**Figure 5.13**) where the observed affect in Gas signalling was neutral (z-score = 0) within the IBD and CD vs control analyses (**Figure 5.7 and Figure 5.10**). It was observed that the DE genes driving the detected effects within the Gas pathway varied between the three IPA analyses (**Table 5.3**), accounting for the fact the pathway was identified in all three analyses.

Table 5.3 | Genes differentially expressed within the Gas signalling pathway

Analysis	Genes differentially expressed
IBD vs Control	*ADCY2, ADCY4, ADORA2B, RAPGEF3, PTGER2, RAP1A*
CD vs Control	*ADCY2, ADCY4, ADORA2B, RAPGEF3, PTGER2, RAP1A, CHRM3, GNG2, GNG12*
UC vs CD	*GPER1, GNG2, RAPGEF3, CREB3L4, ADCY7, CHRM3, GNG12, GNG7*

ADCY7, a gene in which a rare mutation has previously been implicated to give a 2-fold increase in risk for UC [199], was solely observed to be affected within the Gas signalling pathway in UC vs CD analysis (**Table 5.3**). It was observed that the difference in DE genes driving the perturbation within the Gas signalling pathway resulted in different protein complexes and thus arms of the pathway being affected (**Figure 5.14**). Gas signalling is reported to be involved intracellular and second messenger signalling through G-proteins, with functions reported to involve cellular assembly, development, function and maintenance [200-202].

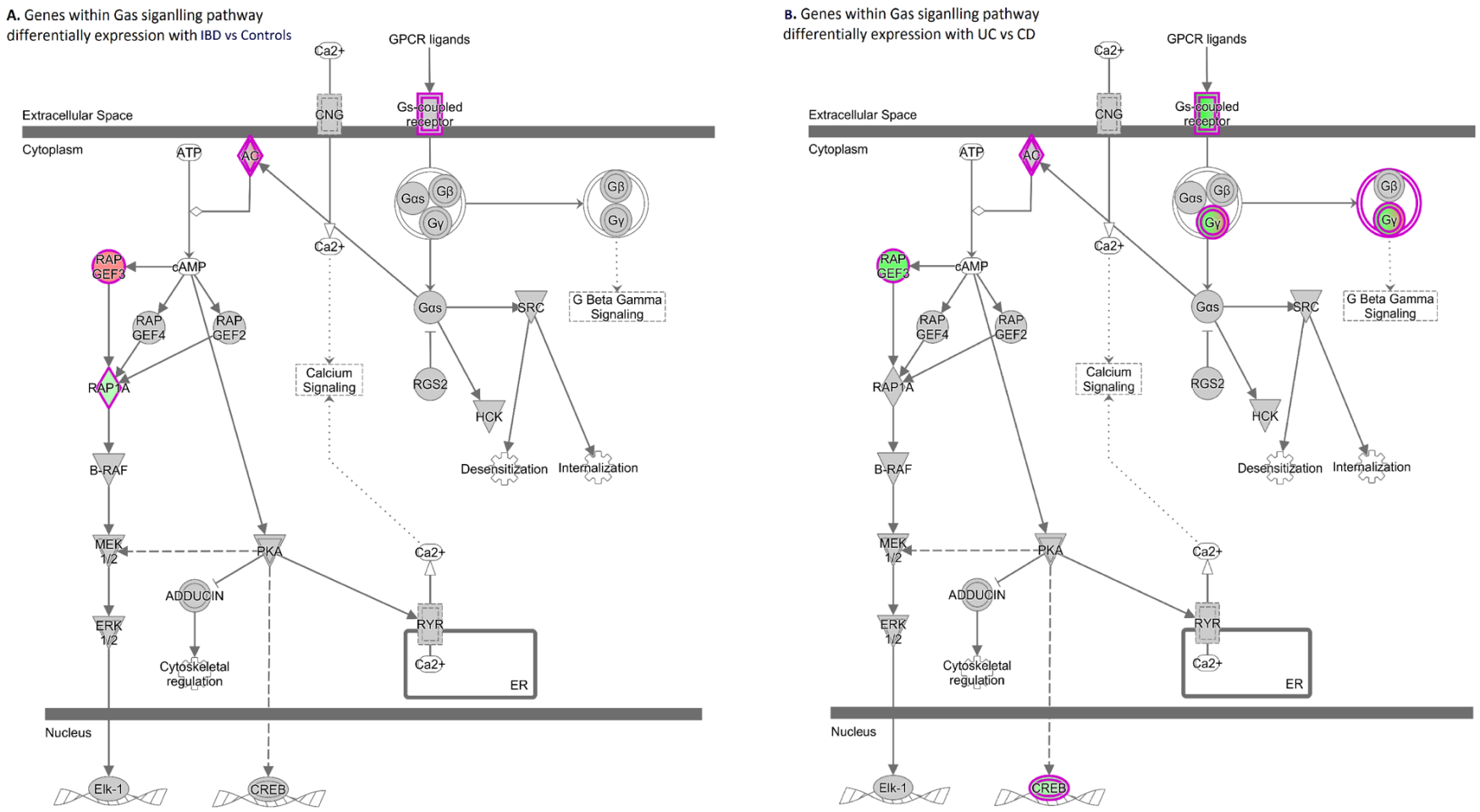


Figure 5.14 | Gas Signalling pathway

Canonical pathway showing Gas signalling pathway with Purple double circles indicating protein complexes containing DE genes and the fill-in colour of the circle indicating direction of effect; downregulation (green) to upregulation (red). With A. showing DE genes for IBD vs controls and B. showing DE genes for UC vs CD. (The pathway analysis was generated through the use of QIAGEN's Ingenuity Pathway Analysis (IPA®, QIAGEN Redwood City, www.qiagen.com/ingenuity).

5.4 Comparison of GSEA and IPA results

Although both forms of analysis – GSEA and IPA – are aimed at generating hypotheses about functional implications of the identified differentially expressed genes, seemingly quite different results were observed. This can be attributed to the form of analyses, e.g. gene set enrichment versus tissue specific canonical pathway analysis. To investigate coherence and increase confidence in the obtained results, the GSEA analysis was repeated for all sub-phenotypes, this time using a canonical pathway dataset as the MSigDB reference dataset. For the IBD vs control analysis IPA identified pathways including Notch signalling, Adaptive Immune system, NK cell pathway and phagosome formation were replicated. Within the CD vs control analysis, the Xenobiotic pathway and metabolism of xenobiotic by cytochrome P450 were identified, replicating previous IPA results. Although no direct matches with IPA were found for the UC vs CD analysis, the CREB pathway was identified which acts downstream of the IPA identified Gas pathway. Furthermore, biological oxidation and drug metabolism pathway were identified related to the IPA identified phosphatase pathway and p450 pathway, respectively.

5.5 Discussion

A pathway analysis was performed to elucidate how genes exhibiting differential expression within the sub-phenotypes of IBD affect biological pathway. Two well-known methods for pathway analysis were employed; Gene Set Enrichment Analysis (GSEA) and Ingenuity Pathway Analysis (IPA). Although both forms of analyses are aimed at generating hypotheses about functional implications of the identified groups of differentially expressed genes, seemingly quite different results were observed. This is most likely due to the form of analyses, e.g. gene set enrichment versus tissue specific canonical pathway analysis. Gene set enrichment compares the presence of differential expressed genes within experimentally generated gene lists reported to be differentially expressed under specific condition, whereas IPA estimated the effect of DE genes on canonical pathways containing genes expressed the gut.

GSEA analysis identified 3,348 gene sets enriched in IBD patients, 59 pathways were enriched in genes solely differentially expressed between CD and controls. Although, 245 pathways suggested to be specific to UC pathogenesis were identified through the UC vs CD DE analysis. The extremely high number of significantly enriched pathways within the GSEA analysis combined with the highly specific nature of the MSigDB, it was decided that the use of IPA was preferable.

Using IPA the IBD vs control DE gene set was enriched for genes involved in 17 canonical molecular pathways. The CD vs control DE analysis was enriched for genes in 28 pathways, 25 of which were unique to this analysis. Furthermore, 21 pathways were identified to be enriched in the list of UC vs CD DE genes, of which 11 were not found in either the IBD or CD analyses suggesting they might be involved in UC pathogenesis.

5.5.1 Gas and G-protein signalling pathways

Gas signalling was the only pathway observed to be perturbed by DE genes within all sub-phenotypes. Gas signalling involves intracellular and second messengers signalling through G-proteins. In addition to Gas signalling, several other pathways involved with G-protein signalling were observed to be perturbed including Gai signalling, cAMP-mediated signalling, GPCR-mediated nutrient sensing in enteroendocrine cells and GPCR-mediated integration of enteroendocrine signalling amplified by an L cell. Different arms of the gas signalling pathway were observed to be effected within CD or UC phenotypes, cAMP signalling was perturbed within the IBD vs control and UC vs CD analyses and perturbations in Gai signalling and both GPCR signalling pathways might be CD specific. G-proteins have well established functions within transmembrane signalling [200] and have been implicated in the regulation of tight junction formation within epithelial cell, with overexpression of Gα subunits increasing epithelium permeability [201,203]. In addition to G-protein involvement in epithelial tight junction formation, *RGS1* (regulator of G protein signalling 1) has been shown to reduce T cell migration to lymphoid-homing

chemokines in the gut [202]. Although, the role of G-proteins has previously been shown to regulate processes important to IBD (transmembrane signalling, tight junction formation, gut permeability and T cell responses in the gut), the data presented in this study was able to add value by providing insight into IBD sub-type specific G-protein pathway perturbations.

5.5.2 Notch signalling pathways

Notch signalling, hematopoiesis from pluripotent stem cells, phosphatidylcholine biosynthesis I, Phosphatidylglycerol biosynthesis II and CDP-diacylglycerol biosynthesis I were all pathways observed to be effected within the IBD vs control analysis but not in the UC vs CD analysis, suggesting they effect IBD pathogenesis. Notch signalling is known to regulate intestinal stem and progenitor cell proliferation and differentiation [204,205]. Notch maintains adult intestinal stem cells to regulate cell fate choice and control epithelial cell homeostasis [195,206,207]. It has been hypothesised that activation of Notch signalling is required for epithelial repair in IBD, with increased Notch activity observed within IBD epithelial cells [196,208]. This is consistent with the observed positive z-score and upregulation of five genes within the Notch signalling pathway in IBD patients. The Notch signalling pathway is comprised of four different receptors, NOTCH1 (Notch 1), NOTCH2 (Notch 3), NOCTH3 and NOTCH4. The two Notch receptors observed to be upregulated within the Notch signalling pathway are Notch 3 and 4, consistent with the observed enrichment by GSEA. Notch 4 has been shown to effect similar target genes to Notch 1, although minor difference in function have been observed with the Notch 4 intracellular domain (ICD) having shown to reduce transforming growth factor beta (TGF-β) induced growth inhibition of mammary epithelial cells [209]. Notch 1 has been shown to modulate mucosal chemokines and cytokine secretion as well as effector T cell responses, regulating protective epithelial pro-inflammatory responses [210]. Notch 3 has been reported to promote neuronal differentiation [211] and inhibit epithelial differentiation in the lung [212]. Furthermore, activation of Notch 3 has been observed to enhance the

generation of regulatory T cells [213]. The observation that Notch 3 and Notch 4 are the affected Notch proteins driving, in part, the observed perturbation of the Notch signalling pathway provided insight into which specific signals of Notch signalling are important in IBD.

5.5.3 Drug metabolism, xenobiotics and nicotine pathways

Several pathways exhibited opposite directions of effect within the CD vs controls and UC vs CD analyses, including Nicotine degradation II and III, two melatonin degradation pathways and Xenobiotic metabolism signalling, suggesting they predominantly affect CD pathogenesis. Smoking and inflammatory bowel disease has been topic of discussion for many years. Smoking has been shown to have opposite effects on the clinical course of UC and CD with smoking being beneficial to clinical remission in UC patients [214] whereas, detrimental effects are reported in CD [215]. With nicotine being an important component in smoking, it is interesting that our analysis showed two nicotine degradation pathways to be perturbed solely within CD patients. Interestingly, a recent study has shown that twice as many CD patients are reported to be active smokers compared to UC patients [216]. By incorporating smoking as a covariate into the here performed differential expression model it was ruled out that the higher percentage of smokers within the CD patient group was responsible for the observed significant perturbation of Nicotine degradation pathways II and III. It has been shown that smoking influences both innate and adaptive immunity with smokers exhibiting reduced cytokine production [217] and altered immunoregulatory T cell ratios [218,219]. Factors like intestinal mobility [220], gut permeability and blood flow have also been investigated [221], but have conflicting or non-conclusive outcomes. Although, a multitude of research studies have been performed, an explanation for the opposing effects of smoking on UC and CD has not yet been found. The analyses performed in this study, which identifies a perturbation of nicotine degradation pathways II and III in CD patients warrants further investigation. Another pathway identified within CD patients, by both IPA and GSEA, was Xenobiotics

metabolism. Xenobiotics are organic compounds to which an organism is exposed that are extrinsic to the normal metabolism and include drugs. The altered expression of xenobiotic metabolism genes is most likely due to drug therapies in the CD patients compared to controls. GSEA identified various additional drug related gene sets with significant enrichment including steroid hormone biosynthesis, glucocorticoid therapy, drug metabolism cytochrome p450 and drug metabolism other enzymes, indicating that drugs and drug metabolism have a significant effect on the transcriptome of IBD patients.

6. Expression Quantitative trait loci (eQTL) analysis in IBD relevant tissue

Genomics is a constantly evolving field; one newly emerging technology is expression quantitative trail loci (eQTL) analysis. eQTL analysis investigates associations between SNPs and changes in gene expression. The abundance of a gene transcript could directly be modified by polymorphisms in regulatory elements, identifying the effect of a SNP on changes in gene expression could provide highly valuable information in complex diseases. Disease associated SNPs located in the coding region of a gene often effect that specific genes, for disease associated SNPs located in non-coding regions their effect is often more complex to identify. eQTL analysis attempts to address this by performing a genome wide linkage analysis between genetic polymorphisms and variation in gene expression. Associations between a SNP and changes gene expression levels can be investigated at a local level (*cis*-eQTL) or at a distant level (*trans*-eQTLs). *Cis*-eQTLs are identified as associations between SNPs and genes located within 1Mb on either side of the SNP. Distant eQTLs (*trans*-eQTLs) are identified as associations between SNPs and genes located beyond 1Mb, potentially even on another chromosome.

The majority of IBD susceptibility SNPs are known to be located in non-coding regions and do not directly alter gene function. It was therefore decided to employ eQTL analysis to attempt and elucidate the effect of IBD susceptibility SNPs on changes in gene expression of nearby genes (*cis*-eQTLs); the study is not powered to detect *trans*-eQTLs. Matrix eQTL was employed to perform a genome wide expression quantitative trait loci analysis, correlating variation in gene expression within the large intestine to underlying genomic variation. Whole transcriptome data was generated from large intestinal tissue for 127 individuals (103 IBD patients and 24 controls) (see Chapter 3). For 121 out of 127 individuals, genome wide SNP genotype data was also generated (see Materials and Methods section 2.2.7). Matrix eQTL calculates significant associations between genotypes/alleles of each of 241,995 input SNPs to changes in gene expression. P-values and False Discovery Rate (FDR) values

are generated via linear regression and the Benjamini Hochberg method, respectively. A significant association was defined as False Discovery Rate (FDR) ≤ 5%.

6.1 Quality control

Cis-eQTLs were identified by testing all SNPs within 1Mb upstream or downstream of the transcription start site of a given gene. Associations between 241,995 indexed SNPs and 17,258 genes were investigated. In total, 2,096 significant *cis*-eQTLs at FDR of 5% (q ≤ 0.05) were identified involving 861 genes. Validity of the observed results was assessed by quantile-quantile (QQ) plot and histogram (**Figure 6.1A-B**).

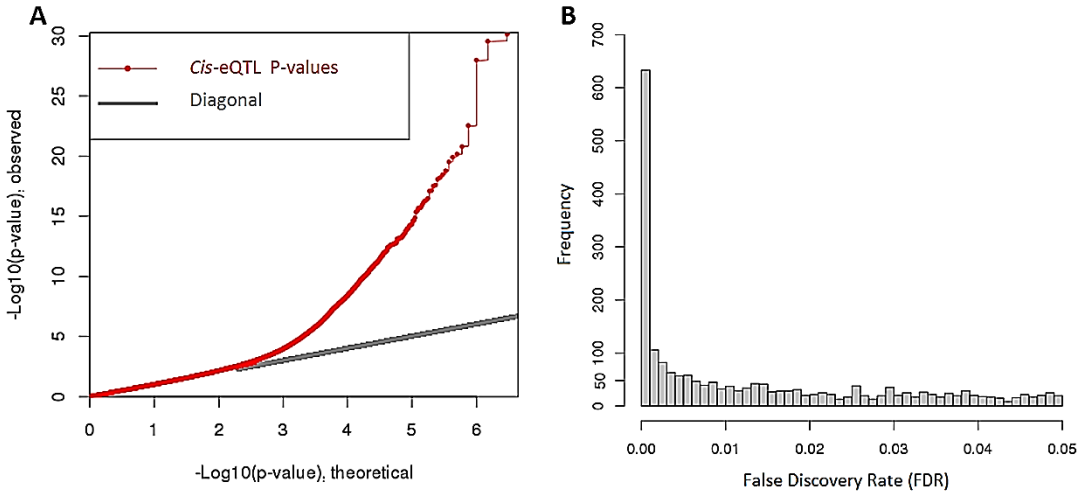


Figure 6.1 | Quality control cis-eQTL results
(A) QQ plot showing theoretical p-values (-log10) against observed p-values (-log10) for *cis*-eQTL (red), with 45 degrees (x=y) line shown in grey. **(B)** Histogram visualising the distribution of significant *cis*-eQTL FDR values, with each bar representing 0.001 FDR. The height of bars indicates the frequency of a given FDR value.

The QQ plot was used to compare genome-wide distribution of the eQTL statistic with the expected null distribution (inflation) (**Figure 6.1A**). Inflation can be introduced through unknown variables such as sample duplication, unknown familiar relationships or technical bias. The genomic inflation factor

(Lambda) was calculated to be 1, indicating the data follows normal chi-square distribution and no inflation was observed. The histogram visualises significance distribution, showing that more than 600 *cis*-eQTLs were highly significant with an FDR of ≤ 0.001 (**Figure 6.1B**).

6.1.1 Multiple correlated eQTL signals per gene

Overall, 2,096 significant eQTLs showed association with 861 genes. Multiple SNPs showing association with a single gene can be caused by high linkage disequilibrium (LD) between the SNPs or indicate multiple individual SNPs signals. For example, the changes in expression of *GSDMB* (Gasdermin B) a gene located on chromosome 17 (IBD locus 17.03), were observed to be significantly associated with 17 SNPs. LD was evaluated using locus-zoom, based on 1000 Genomes data, for the index SNP rs10852936 and all SNPs located within 1Mb of *GSDMB* (**Figure 6.2**).

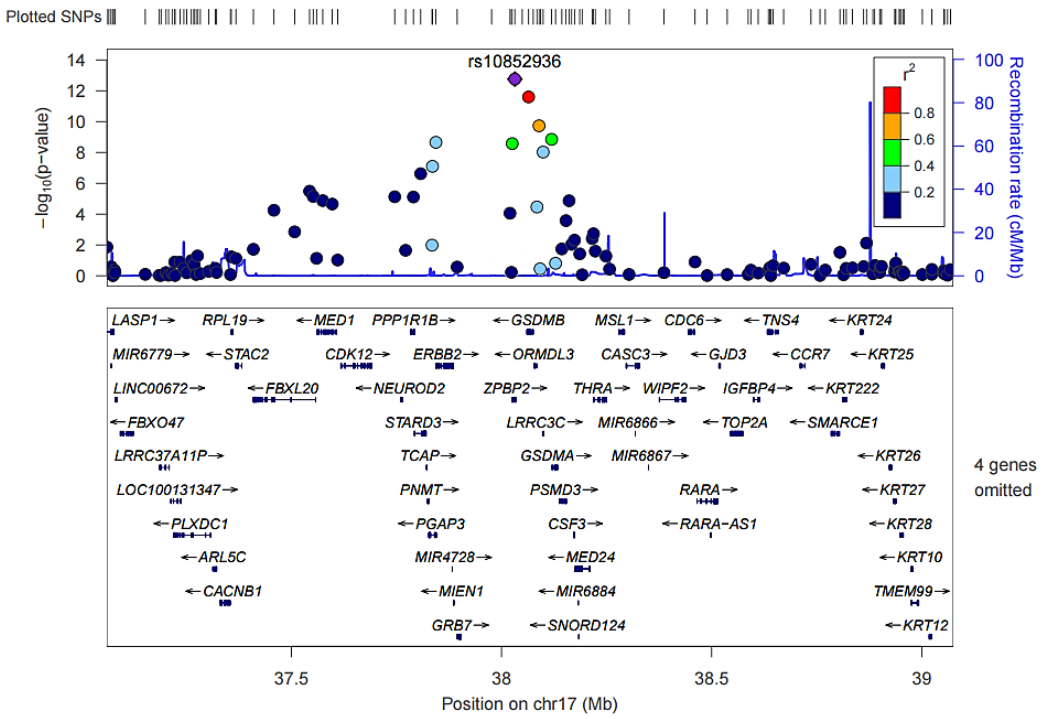


Figure 6.2 | GSDMB expression quantitative trait locus (eQTL) effects around SNP rs10852936

Locus-zoom plot of the *GSDMB* eQTL effect using hg19/1000 Genomes data for SNP location and linkage disequilibrium scores. Showing linkage disequilibrium (LD) scores (r^2) between the index SNP, rs10852936, and all SNPs located within 1MB of *GSDMB*. With r^2 ranging from 1, high LD (red), through to 0, low LD (dark blue). Genes name in box indicate all genes present within the genomic region where the SNPs are located. (plot generated with Locus-Zoom http://locuszoom.org/).

LD between rs10852936 and the 16 additional significant eQTL SNPs was observed to range from $r^2>0.8$ to $r^2<0.2$, with both LD and eQTL significance decreasing as the distance from the index SNP increased (**Figure 6.2**). Similar results regarding LD were observed for the other genes containing multiple eQTLs. Although LD calculations can give an indication of the level of correlation between SNPs, in depth fine-mapping would be required to determine with certainty if there is more than one independent SNP driving the observed eQTL signals.

6.1.2 Genotype coverage of IBD loci locations

To assess coverage of the 224 known IBD susceptibility loci by the chosen genome-wide SNP genotyping array, Dr Ken Hanscombe a research associate

within the statistical genetics unit, determined the level of LD between the 502 index SNPs from the IBD associated loci [92,107,108,112] and the 241,995 genotyped SNPs (see Materials and Methods section 2.2.9.6.4). It was observed that 255 out of 502 IBD SNPs were captured by our data (83 genotyped directly, 172 in high LD at $r^2 \geq 0.8$) corresponding to 118 out of the 224 known IBD susceptibility loci.

6.2 *Cis*-eQTLs within known IBD susceptibility loci

To further prioritise the intestinal *cis*-eQTL findings in IBD, it was considered which of the significant *cis*-eQTL genes were located within 500kb of the 224 known IBD susceptibility loci. It was observed that 126 genes with intestinal *cis*-eQTLs were located within one of 76 IBD susceptibility loci (**Figure 6.3**). Five of these genes had also been highlighted as significant in our intestinal differential expression data. Furthermore, 25 of these genes had been previously prioritised by earlier bioinformatic studies investigating gene and protein networks across all IBD GWAS loci [107,150] (from now on these will be called "previously prioritised genes") (See Chapter 4.3.1). Chromosome 6 was observed to contain the highest number of *cis*-eQTLs (19) at high significance (**Figure 6.3**). Chromosome 6 contains 19 IBD loci, 2 of which encompass the human leukocyte antigen (HLA) class I and II regions encoding the major histocompatibility complex (MHC) proteins in humans. *Cis*-eQTL signals were confirmed within 24 previously prioritised genes (green dots) and 4 genes shown to exhibit differences in expression between IBD and controls (yellow dots), suggesting that these genes may be important in IBD pathogenesis (**Figure 6.3**). The gene *FAM49B* (Family With Sequence Similarity 49, member B) located on chromosome 8 (chr8:130077267-131124661) was the only gene identified as a significant *cis*-eQTL which was also differentially expressed in intestinal tissues in IBD and previously prioritised by other studies (indicated by a red dot in **Figure 6.3**). The remaining 97 genes representing significant *cis*-eQTLs in known IBD loci have not previously been implicated in connection with IBD pathology or etiology.

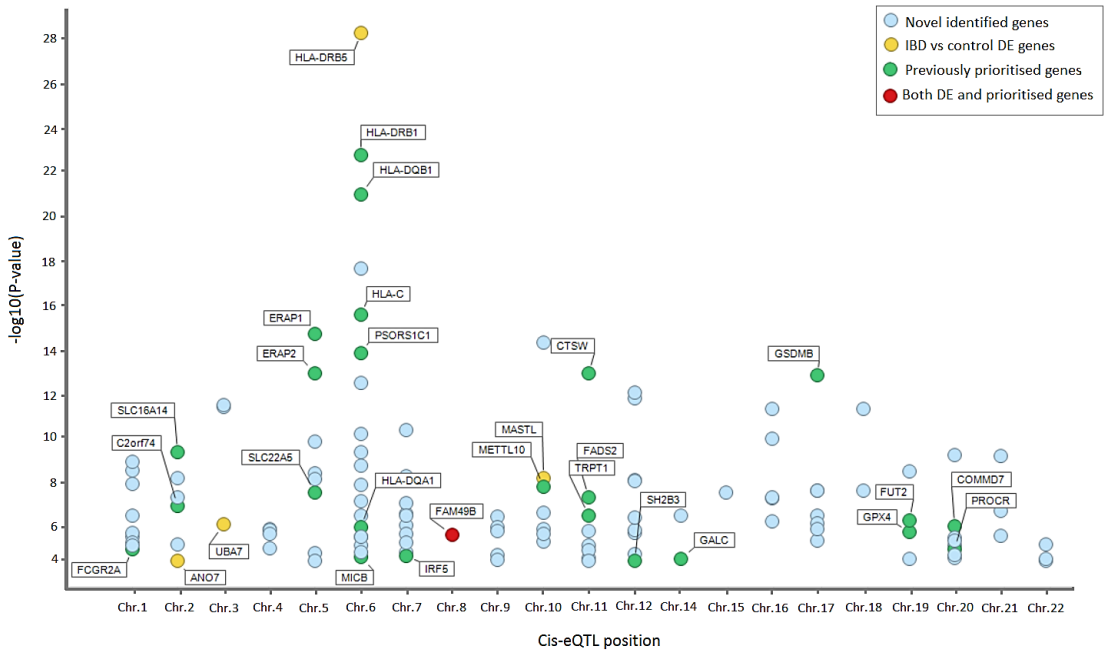


Figure 6.3 | Significant Cis-eQTLs within IBD loci
Significant identified cis-eQTLs per chromosome, each dot representing a gene located within the genomic boundaries of a known IBD susceptibility locus plotted by levels of significance (-log10(p-value)). With previously prioritised genes shown in green, differentially expressed genes between IBD cases and controls in yellow, genes both DE and prioritised in red and novel identified genes in light blue.

6.3 Intestinal *cis*-eQTLs at known IBD susceptibility SNPs

Overall, 126 genes in *cis*-eQTL were identified to be located within an IBD susceptibility loci. To further explore our eQTL results in relation to IBD, the eQTL SNPs were compared to a list of 6,931 IBD associated SNPs compiled from recent GWAS studies [92,107,108,112] using LDlink [222]. It was established that 24 genes (out of 126) in *cis*-eQTL were associated with SNPs in high LD ($r^2 \geq 0.7$) with IBD susceptibility SNPs. Furthermore, 8 out of 126 *cis*-eQTLs SNPs were a direct match with an IBD susceptibility SNP, of which 3 were GWAS index SNPs (**Table 6.1**).

Table 6.1 | Cis-eQTLs associated with IBD susceptibility SNPs

Gene Name	IBD locus	Locus location (Mb)	SNP	P-value	FDR	Beta Score
CTSW	11.06	65.1–66.2	rs568617*	1.5E-13	8.6E-09	-0.96
GPX4	19.01	0.6–1.6	rs4807569	1.7E-06	5.5E-03	0.58
GSDMB	17.03	37.4–38.6	rs10852936	1.7E-13	9.7E-09	-0.79
UQCR11	19.01	0.6–1.6	rs4807569	2.9E-05	4.4E-02	0.345
WDR6	3.03	47.9–51.6	rs11715581*	3.1E-06	8.4E-03	0.39
SFMBT1	3.04	52.5–53.6	rs9847710*	3.9E-12	1.2E-07	-0.85
RGS14	5.17	176.3–177.3	rs4976646*	4.1E-09	4.1E-05	-0.53
ERAP2	5.08	95.7–96.9	rs7719705	8.6E-09	1.5E-13	0.91

* IBD index SNPs

ERAP2 (Endoplasmic Reticulum Aminopeptidase 2), *CTSW* (Cathepsin W), *GSDMB* and *GPX4* (Glutathione Peroxidase 4) were previously prioritised genes in IBD (see Chapter 4.3.1), while the remaining four genes were not included on the previously prioritised gene list. *WDR6* (WD Repeat Domain 6), *RGS14* (Regulator of G-protein Signalling 14) and *UQCR11* (Ubiquinol-Cytochrome C Reductase, Complex III subunit XI) map to IBD regions already containing prioritised genes, while *SFMBT1* (Ubiquinol-Cytochrome C Reductase, Complex III subunit XI) maps to an IBD region on chromosome 3 (chr3:52478418-53642980) without any previously prioritised genes (**Figure 6.4**).

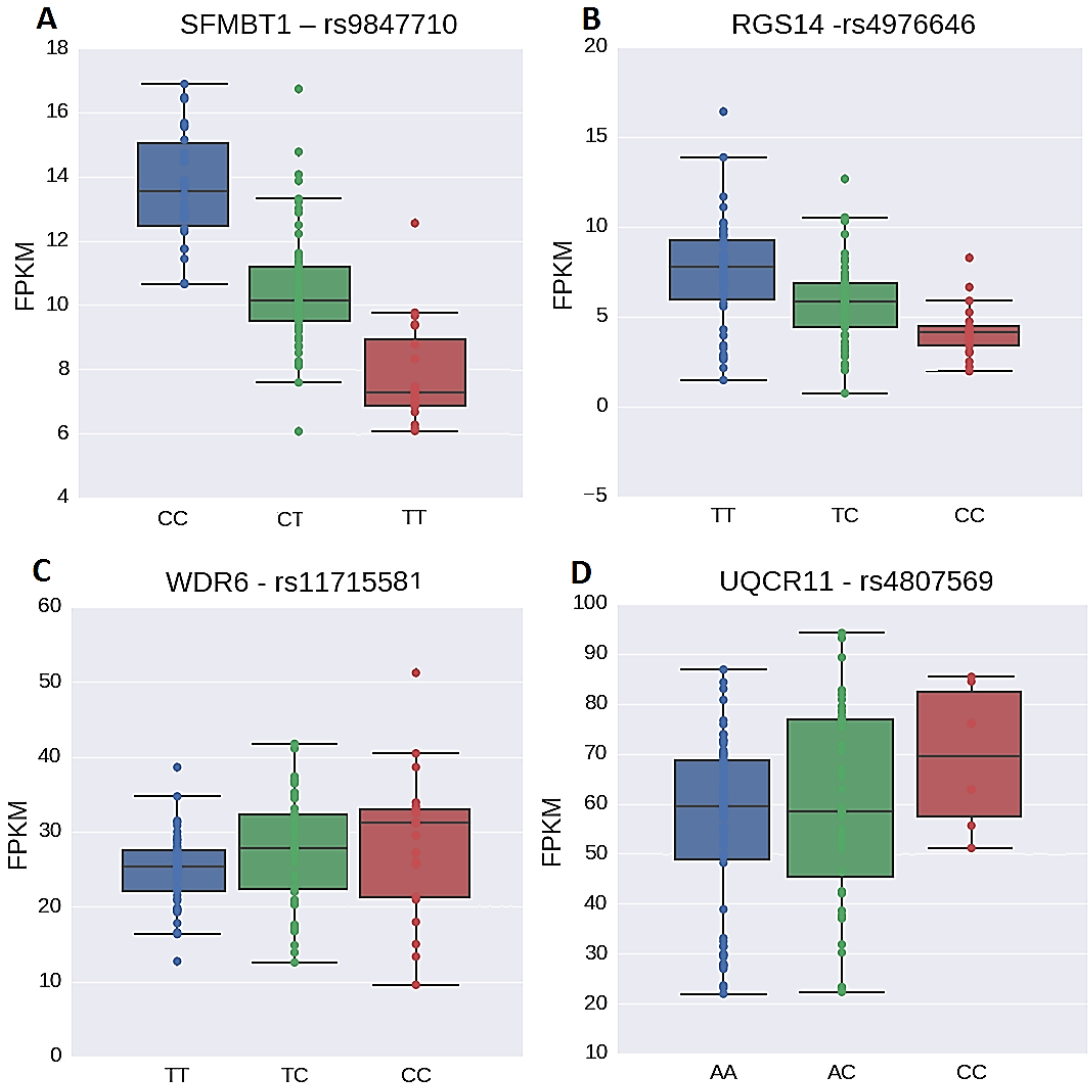


Figure 6.4 | cis-eQTL in novel genes associated with IBD susceptibility SNPs
Levels of expression in FPKM (Fragments Per Kilobase of exon per Million fragments mapped) associated with genotypes for IBD susceptibility SNPs for *SFMBT1* (**A**), *RGS14* (**B**), *WDR6* (**C**) and *UQCR11* (**D**). Blue indicates homozygotes for the major allele, green heterozygote and red homozygotes for the minor allele. False discovery rate (FDR) following Benjamini Hochberg correction for multiple testing are (**A**) 1.2×10^{-7}, (**B**) 4.1×10^{-5}, (**C**) 8.0×10^{-3} and (**D**) 4.0×10^{-2}.

SFMBT1 expression was reduced 2-fold, from 13.6 FPKM within CC homozygotes to 7.8 FPKM in TT homozygotes ($q = 1.2 \times 10^{-7}$) (**Figure 6.4A**). SFMBT1 has been reported to function as a histone-binding protein which mediates recruitment of corepressor complexes to target genes [223]. *RGS14* expression showed a near 2-fold reduction, from 7.5 FPKM to 4 FPKM, in homozygotes with major T allele compared to homozygotes with the minor C

allele ($q = 4.1x10^{-5}$) (**Figure 6.4B**), and is reported to regulate G protein-coupled receptor signalling cascades influencing cell division and stress resistance [224]. *WDR6* and *UQCR11* showed increased expression within heterozygotes and homozygotes for the minor C allele, $q = 8.0x10^{-3}$ and $4.0x10^{-2}$, respectively (**Figure 6.4C-D**). *WDR6* expression increased slightly, with approximately 25 FPKM in homozygotes for the major T allele and around 32 FPKM observed in samples with minor CC allele (**Figure 6.4C**). WDR6 has been suggested to interact with STK11 (serine/threonine kinase 11) and induce cell growth arrest [225]. *UQCR11* shows the smallest change in expression out of the 4 genes visualised in figure 6.4, although still significant at $q = 4.0e^{-2}$, with mean expression increasing from 55.7 FPKM in homozygotes for the major A allele to 79.2 FPKM in homozygotes for the minor allele C (**Figure 6.4D**). UQCR11 has been reported to be part of a protein complex involved in the mitochondrial respiratory chain [226].

6.4 *Cis*-eQTLs associated with previously prioritised genes in IBD

Out of 126 genes in significant *cis*-eQTL and located within an IBD locus, 29 were previously reported to be potentially involved in IBD pathogenesis. Of these, 25 were previously prioritised genes, and four genes: *ANO7, UBA7, HLA-DRB5* and *MAST* were shown to exhibit differential expression between IBD cases and controls (see Chapter 4.2.1). *FAM49B* was the only gene to be previously prioritised, exhibit differential expression and be in a *cis*-eQTL. Our identification of these 29 genes in the eQTL analysis strengthens that hypothesis (**Table 6.2**).

Table 6.2 | cis-eQTLs associated with genes previously prioritised to be involved in IBD pathogenesis

Gene Name	IBD loci	Loci location	Top eQTL SNP	P-value	FDR	Beta score
FCGR2A	1.18	160.9 – 161.9	rs12116744	1.1E-05	2.2E-02	0.49
C2orf74	2.05	60.7 – 61.7	rs720201	1.2E-07	6.3E-04	0.73
SLC16A14	2.20	230.6 – 321.7	rs1124534	4.7E-10	6.7E-06	1.07
ANO7	2.23	242 - 243	rs13411510	3.3E-05	4.8E-02	-0.57
UBA7	3.03	48 – 51.6	rs6446298	8.2E-07	3.0E-03	0.52
ERAP1	5.08	95.7 – 96.9	rs27045	2.5E-15	2.6E-10	0.88
ERAP2	5.08	95.7 – 96.9	rs7719705	1.5E-13	8.6E-09	0.91
SLC22A5	5.09	129.2 – 132.3	rs272885	3.2E-08	2.1E-04	0.56
HLA-C	6.07	30.74 – 31.8	rs2524074	3.6E-16	4.4E-11	0.97
PSORS1C1	6.07	30.7 – 31.8	rs3094217	1.8E-14	1.4E-09	-0.88
MICB	6.07	30.7 – 31.8	rs2516408	2.2E-05	3.7E-02	0.57
HLA-DRB5	6.08	32.1 – 33.1	rs9270986	1.0E-28	1.0E-22	1.78
HLA-DRB1	6.08	32.1 – 33.1	rs9270986	2.9E-23	2.2E-17	1.62
HLA-DQB1	6.08	32.1 – 33.1	rs3135006	1.6E-21	9.5E-16	1.35
HLA-DQA1	6.08	32.1 – 33.1	rs3135006	1.1E-06	3.6E-03	0.75
IRF5	7.13	128.1 – 129.1	rs3757385	2.1E-05	3.4E-02	-0.42
FAM49B	8.06	130.1 – 131.1	rs13340584	2.3E-06	6.7E-03	0.53
MASTL	10.02	26.6 – 27.7	rs1981296	6.9E-09	6.2E-05	-0.63
METTL10	10.13	125.8 – 127.1	rs1055256	1.6E-08	1.2E-04	0.49
FADS2	11.04	61 – 62.1	rs174593	5.3E-08	3.1E-04	0.74
TRPT1	11.05	63.6 – 64.7	rs11603384	3.4E-07	1.5E-03	0.69
CTSW	11.06	65.1 – 66.2	rs568617	1.5E-13	8.7E-09	-0.96
SH2B3	12.06	102.9 – 114.3	rs11065934	3.3E-05	4.8E-02	0.69
GALC	14.03	87.9 – 89.1	rs10483987	2.7E-05	4.2E-02	-0.91
GSDMB	17.03	37.4 – 38.6	rs10852936	1.7E-13	9.7E-09	-0.80
GPX4	19.01	0.6 – 1.6	rs4807569	1.7E-06	5.3E-03	0.58
FUT2	19.05	48.7 – 49.8	rs281380	5.3E-07	2.1E-03	0.48
COMMD7	20.02	30.2 – 31.9	rs12480157	9.5E-07	3.4E-03	0.57
PROCR	20.03	33.3 – 34.4	rs2295888	1.1E-05	2.1E-02	-0.88

Changes in *FAM49B* expression were significantly associated with SNP rs13340584, FDR 0.006 (**Figure 6.5**). Homozygotes for the minor T allele showed a 2-fold increase of *FAM49B* expression (**Figure 6.5**). Limited functional knowledge is available for *FAM49B*; it is reported to be highly expressed in patients with multiple sclerosis and non-small cell lung cancer

tissues [227,228]. Furthermore, it is suggested to be the source of the antigenic peptide presented by Qa-1^b cells in absence of ERAAP [229].

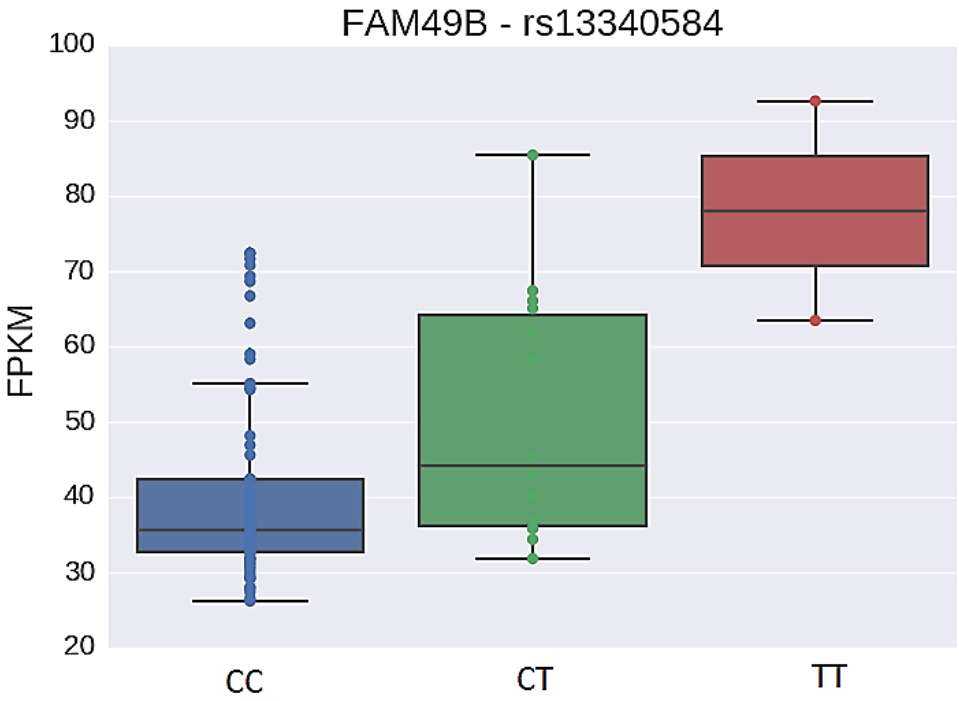


Figure 6.5 | Changes in FAM49B expression associated with rs13340584
Levels of FAM49B expression in FPKM (Fragments Per Kilobase of exon per Million fragments mapped) associated with genotypes for rs13340584, with C being the major and T being the minor allele. Blue are homozygotes for major C allele, green are heterozygote and red are homozygotes for minor T allele. n=85 and false discovery rate (FDR) following Benjamini Hochberg correction = 0.006.

6.5 Novel *cis*-eQTLs located within IBD susceptibly loci

In addition to confirming 29 previously prioritised genes, 97 of the significant *cis*-eQTL findings located within known IBD loci were associated with expression of genes not previously prioritised in IBD. In the earlier chapter (see Chapter 4.3.2) 45 IBD loci were identified for which the previous GWAS studies and bioinformatics analyses had failed to identify any plausible candidate genes. The data generated here has successfully identified 15 intestinal *cis*-eQTLs (**Table 6.3**), genes in these 45 IBD loci. The 15 genes identified were

observed to fall within 12 out of 45 such IBD susceptibility loci, with locus 7.10 (chr7:99901433-100933794) containing two eQTLs: *TRIP6* (Thyroid Hormone Receptor Interactor 6) associated with rs7784933 and *FIS1* (Fission, Mitochondrial 1) associated with rs6979122 (**Table 6.3**). Additionally, locus 21.03 (chr21:35220000-36240000) contained three genes in eQTL: *WRB* (Tryptophan Rich Basic Protein) associated with rs2836995, *PSMG1* (Proteasome Assembly Chaperone 1) associated with rs2297256 and *LCA5L* (Leber Congenital Ameurosis 5-Like) associated with rs2836999, FDR = $1.3x10^{-3}$, $7.39x10^{-3}$ and $4.63x10^{-4}$, respectively (**Table 6.3**). The remaining 10 loci contained one *cis*-eQTL each (**Table 6.3**).

Table 6.3 | Novel *cis*-eQTLs located within IBD loci previously lacking putative candidate functional genes in IBD

Gene Name	IBD loci	loci location (Mb)	Top eQTL SNP	P-value	FDR	Beta Score
BTBD8	1.10	92.1-93.1	rs12129878	2.9E-06	8.1E-03	0.90
HAAO	2.04	43.0-44.4	rs3821349	6.6E-06	1.5E-02	-0.52
SFMBT1	3.04	52.5-53.6	rs9847710	3.9E-12	1.2E-07	-0.85
NQO2	6.02	2.9-3.9	rs2070998	3.2E-07	1.4E-03	-0.63
ECHDC1	6.13	126.9-128.0	rs9398840	1.3E-08	1.1E-04	0.57
NUP43	6.17	149.1-150-1	rs12529698	3.0E-06	8.3E-03	-0.56
FIS1	7.10	99.9-100.9	rs6979122	2.2E-06	6.3E-03	-0.70
TRIP6	7.10	99.9-100.9	rs7784933	5.9E-09	5.5E-05	0.66
MET	7.12	116.4-117.4	rs6977929	1.4E-05	2.3E-02	0.73
SPATA6L	9.01	4.5-5.5	rs6476893	3.8E-07	1.6E-03	0.72
HNRNPA1P70	12.05	68.0-69.0	rs1468487	1.5E-06	4.7E-03	-0.59
LMAN1	18.03	56.4-57.4	rs1899894	2.4E-08	1.7E-04	0.55
LCA5L	21.03	40.0-41.0	rs2836999	8.4E-08	4.6E-04	0.65
PSMG1	21.03	40.0-41.0	rs2297256	2.6E-06	7.4E-03	0.59
WRB	21.03	40.0-41.0	rs2836995	7.5E-10	1.4E-03	0.72

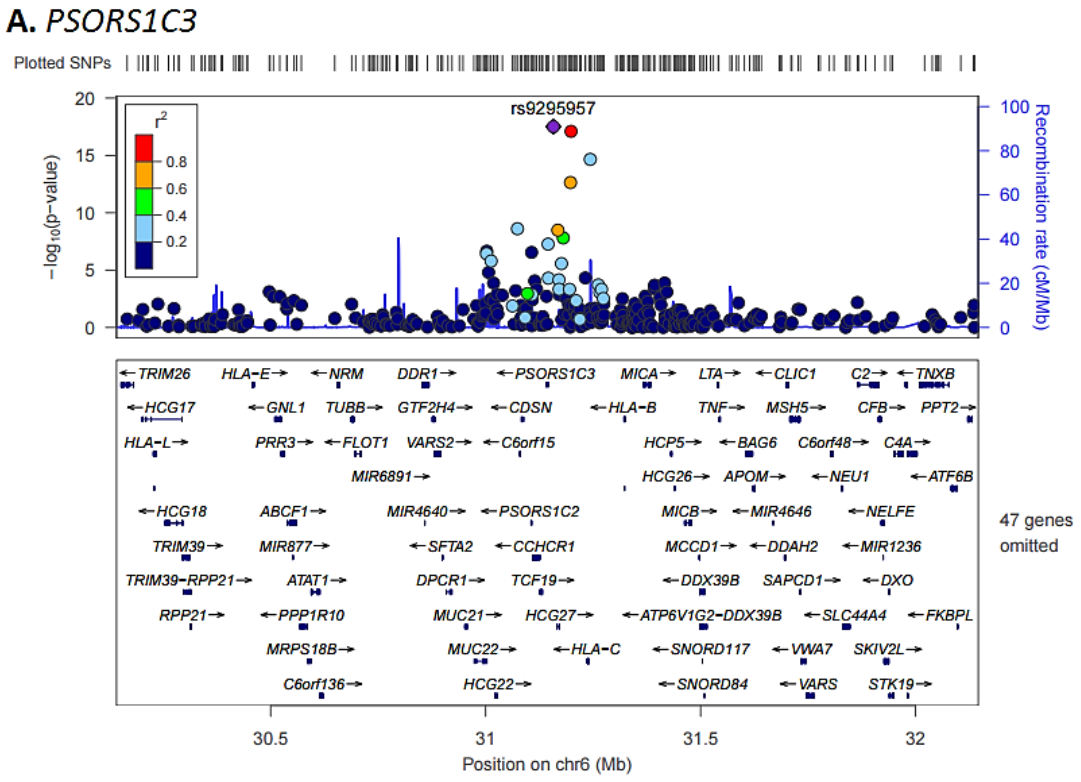
Six out of 15 genes were observed to exhibit a negative beta score, indicating that the minor allele of the SNP in that locus was associated with increased gene expression (**Table 6.3**). The 9 genes identified with a positive beta score showed that the minor allele of the SNP was associated with a decreased gene expression (**Table 6.3**). *FIS1*, *TRIP6*, *MET* and *PSMG1* were reported to contribute to processes previously implicated in IBD development such as

apoptosis [230], inflammatory responses [231] and cell proliferation, migration and survival [232-235] (**Table 6.4**). Furthermore, *SFMBT1* was reported to regulate repression of genes required for development and differentiation by histone-binding [236] (**Table 6.4**). Limited functional knowledge has been reported on the remaining novel identified genes, which increases the difficulty to elucidate their potential involvement in IBD pathogenesis, and might also suggest why they have not previously been thought to be important in this disease (**Table 6.4**).

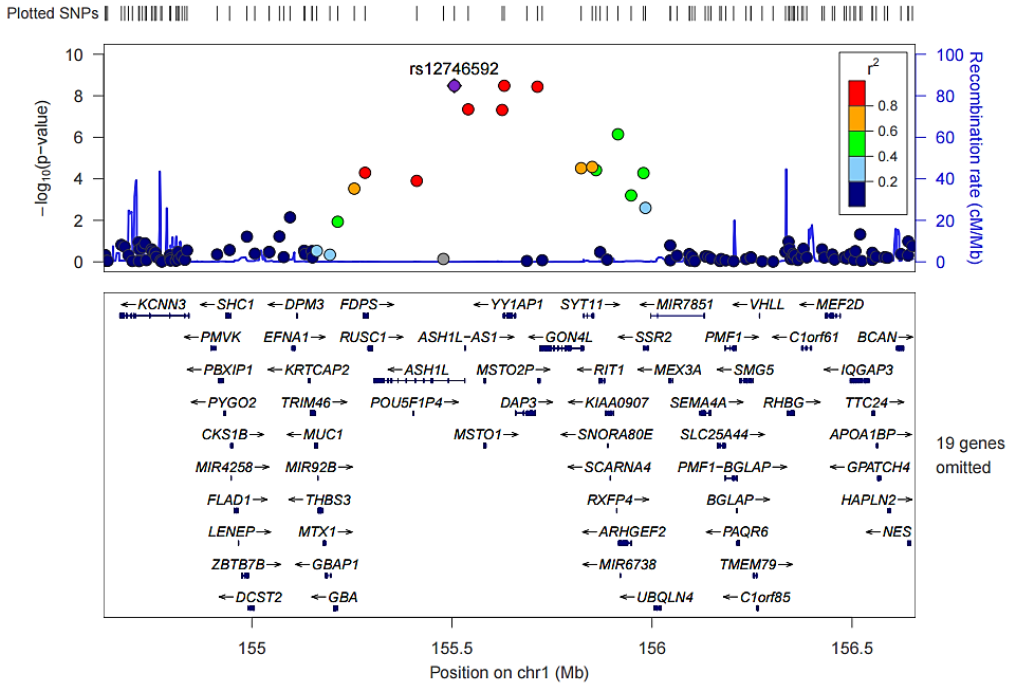
Table 6.4 | Functional information on novel *cis*-eQTLs

Gene	Function information
BTBD8	Contains double BTB/POZ domain, speculated to be important in development
ECHDC1	Suggested to be involved in metabolism of decarboxylases ethylmalonyl-CoA decarboxylase, a toxic metabolite
FIS1	Can induce cytochrome c release from mitochondrion, leading to apoptosis
HAAO	Catalyses the synthesis of quinolinic acid (QUIN), increased QUIN might be involved in inflammatory responses.
HNRNPA1P70	Reported as a pseudogene, no functional knowledge known
LCA5L	Localizes in the nucleus, no functional knowledge known
LMAN1	Cargo receptor for glycoprotein transport
MET	Receptor for hepatocyte growth factor, regulates processes including proliferation, scattering, morphogenesis and survival
NQO2	Serves as a quinone reductase, mutations have been related to neurodegenerative disease and cancers.
NUP43	Enables bi-directions transport macromolecules between cytoplasm and nucleus
PSMG1	Promotes assembly of proteasome subunits, suggested to regulate cell proliferation
SFMBT1	Histone-binding protein, mediates recruitment of corepressor complexes to target genes
SPATA6L	Protein coding gene, no function knowledge known
TRIP6	Involved in lysophosphatidic acid-induced cell adhesion and migration. Additionally, acts as a transcriptional coactivator for NF-kappa-B and JUN
WRB	Localizes in the nucleus, no functional knowledge known

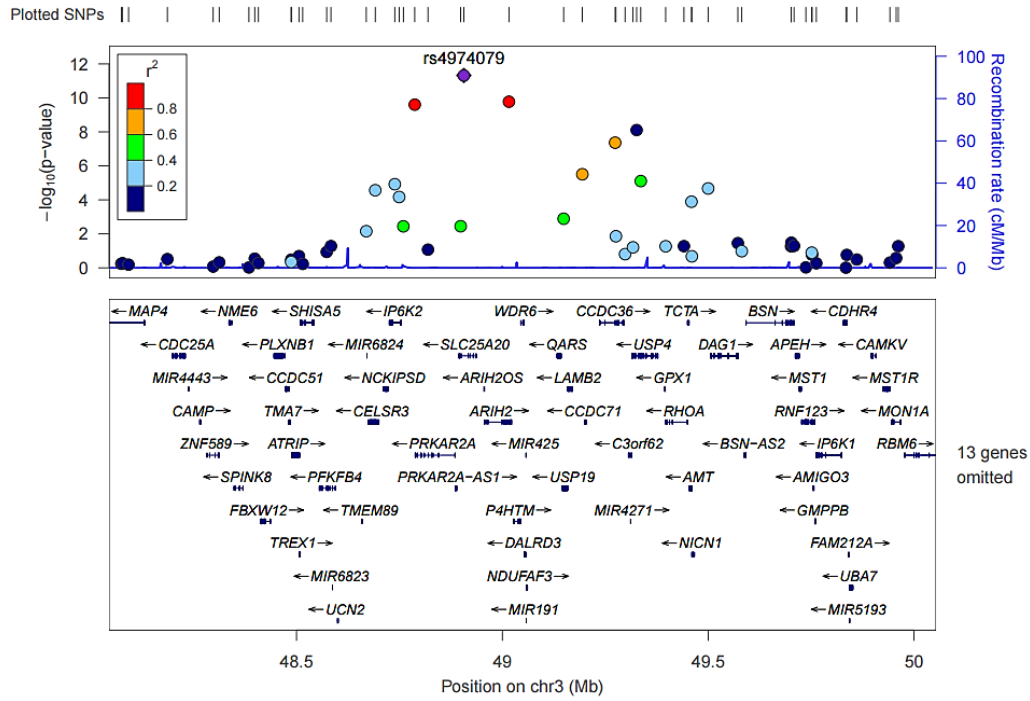
The remaining 82 novel intestinal *cis*-eQTLs which were located within 179 IBD susceptibility loci which contained previously prioritised genes. Most notably, *PSORS1C3* (Psoriasis Susceptibility 1 Candidate 3) with rs9295957 ($q = 6.63x10^{-13}$), *YY1AP1* (YY1 Associated Protein 1) with rs12746592 ($q = 3.35x10^{-5}$), *WDR6* (WD Repeat Domain 6) with rs4974079 ($q = 1.4x10^{-7}$) and *LRRC23* (Leucine Rich Repeat Containing 23) with rs1007924 at $q = 6.65x10^{-8}$ (**Figure 6.6**).



B. *YY1AP1*



C. *WDR6*



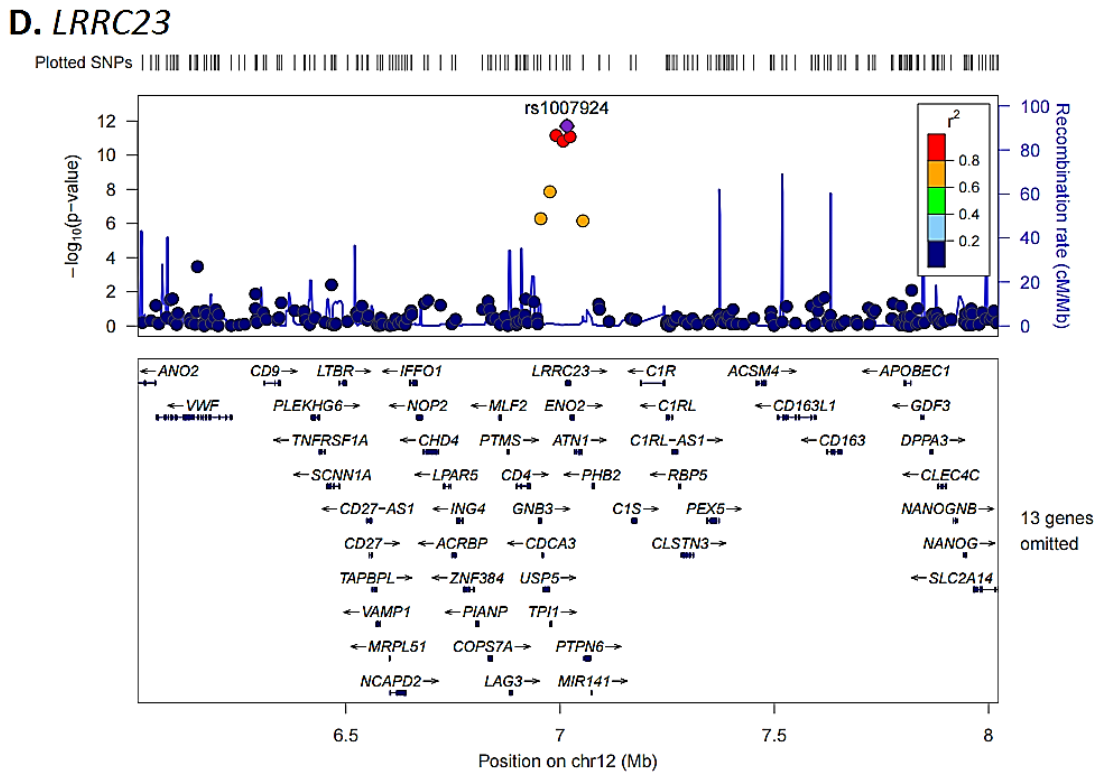


Figure 6.6 | Expression quantitative trait locus (eQTL) effects around top SNPs
Locus-zoom plots of the *PSORS1C3* (**A**), *YY1AP1* (**B**), *WDR6* (**C**) and *LRRC23* (**D**) eQTL effect using hg19/1000 Genomes data for SNP location and linkage disequilibrium scores. Showing linkage disequilibrium (LD) scores (r^2) between the index SNPs rs9295957 (**A**), rs12746592 (**B**), rs4974079 (**C**) and rs1007924 (**D**) and neighbouring SNPs also in eQTL with the gene of interest. With r^2 ranging from 1, high LD (red), through to 0, low LD (dark blue). Genes name in the box below the plot indicate all genes present within the genomic region where the SNPs are located. (plots generated with Locus-Zoom http://locuszoom.org/).

Genotyped SNP coverage of the 2 Mb genomic region surrounding the index SNP was observed to be scarce for *YY1AP1* (**Figure 6.6B**) and *WDR6* (**Figure 6.6C**), while neighbouring SNPs were more abundant for *PSORS1C3* (**Figure 6.6A**) and *LRRC23* (**Figure 6.6D**). For all four genes, the strongest LD (r^2) was observed to affect SNPs with significant *cis*-eQTL scores indicating the observed multiple *cis*-eQTL signals were due to LD and not individual *cis*-eQTLs. *PSORS1C3*, a long non-coding RNA (lncRNA), it has been associated with psoriasis and rheumatoid arthritis (RA) but very little functional information is known [237,238]. Similarly, there is very little functional information about *YY1AP1* and *LRRC23*. YY1AP1 is reported to be a co-activator of transcription factor YY1, and suggested to be involved in cell cycle regulation [239]. LRRC23 is

known to interact with CD28 protein in a pathway regulating the development of regulatory T cells (Tregs) [240]. *WDR6* has been suggested to regulate cell growth arrest and amino acid starvation-induced autophagy [241,242]. Furthermore, chromosome 6, containing the HLA region, showed several novel significant *cis*-eQTL genes including *MICA* (MHC Class Polypeptide-Related Sequence A), *CDSN* (Corneodesmosin) and four HLA genes: *HLA-DQA2*, *HLA-DPB1*, *HLA-DPA1* and *HLA-DQB2*.

6.6 GTEX comparison

Out of 126 *cis*-eQTLs 32 were identified to be associated - or in high LD ($r^2 \geq 0.7$) - with an IBD susceptibility SNP, prioritising these 32 genes as candidates for involvement in IBD pathogenesis. The Genotype-Tissue Expression Project (GTEx), containing eQTLs data across 44 tissues, reported 23 out of the 32 *cis*-eQTLs associated with IBD risk SNPs to have been previously observed within colonic tissue. Out of 32, 9 *cis*-eQTLs are novel within colonic tissue, of which 5 eQTLs associated with genes *C2orf74, NIPSNAP1*, *CEP192*, *PROCR* and *GALC*, have been previously reported to be present in other tissues [92,127]. The remaining four *cis*-eQTLs associated with changes in gene expression of *BORCS7* (Bloc-1 Related Complex Subunit 7), *MAP4K2* (Mitogen-Activated Protein Kinase Kinase Kinase Kinase 2), *UQCR11* (Ubiquinol-Cytochrome C Reductase, Complex III Subunit XI) and *IGLVI-70* (Immunoglobulin Lambda Variable (I)-70) are novel discoveries. The functional processes involving MAP4K2 and IGLVI-70 are both highly relevant to a healthy immune response with MAP4K2 acting as an upstream activator of stress-activated protein kinase/c-Jun N terminal kinase (JNK) signalling pathway [243] and IGLVI-70 being a variable of the lambda chain of immunoglobulin molecules, which are involved in antigen recognition [244]. Although the functional influence of *BORCS7* and *UQCR11* on IBD might not immediately be obvious, the presence of significant IBD risk SNP associated *cis*-eQTLs within colonic tissue warrants further investigation of these genes.

6.7 Discussion

To elucidate the effect of IBD risk SNPs upon gene expression levels within the large intestine expression quantitative trait (eQTL) analysis was performed. Although, eQTL analysis is a powerful tool and it can be used to generate invaluable insights into the effect of disease associated SNPs on the abundance of transcripts, the input files (e.g. the genotype and gene expression data) determine the quality and reliability of the eQTL results. The gene expression data was generated through whole RNA sequencing and has been shown to be of high quality (see Chapter 3). The genome-wide SNP genotype data was generated through the use of two different Infinium arrays: the Human Core or the Human Core Exome Array (see Material and Methods 2.2.7). The two arrays have a large overlap in SNPs, but it should be considered that the use of two different arrays leads to lower power in SNPs only represented on one of the arrays. This was, in part, addressed by applying a filter to only include SNPs genotyped in > 50% of the samples. Furthermore, coverage by the genotype arrays of disease associated SNPs and disease associated loci, plays a major role in the ability to assess eQTLs associated with the disease of interest. The genotype arrays used here covered 118 out of 224 known IBD disease associated loci; meaning we could not assess the presence of eQTLs for approximately 47% of disease associated loci. This should be addressed in future works, either by inputting the data to increase SNP coverage or by re-genotyping the samples on Infinium arrays offering better coverage of the known IBD susceptibility loci.

Overall, 861 significant (FDR<0.05) *cis*-eQTLS were identified of which 126 correlated with genes located within the genomic boundaries of IBD susceptibility loci. Moreover, 32 out of 126 *cis*-eQTLs were identified to be associated - or in high LD ($r^2 \geq 0.7$) - with an IBD risk SNP, providing compelling evidence for these 32 genes to be involved in IBD pathogenesis. The Genotype-Tissue Expression Project (GTEx) [127], containing eQTLs data across 44 tissues, was employed to confirm the 32 colonic *cis*-eQTLs associated with IBD risk SNPS identified within this study.

6.7.1 *cis*-eQTLs implicated in IBD

From the 861 c*is*-eQTLs identified within large intestinal tissue from IBD patients and controls, 32 c*is*-eQTLs were shown to be associations between IBD risk SNPs and genes located within known IBD susceptibility loci. Employing GTEx, it was confirmed that 23 out of 32 *cis*-eQTLs have been previously reported within colonic tissue. The replication of these previously descripted results confirms the robustness and validity of the results produces in the here performed analysis. Out of 23 replicated *cis-e*QTLs, 10 associated genes were previously prioritised in IBD, including *ERAP1*, *ERAP2, GSDMB* and *FADS2*. ERAP1, ERAP2 (Endoplasmic reticulum aminopeptidase 1 and 2) are involved in the processing of peptides prior to antigen presentation through the major histocompatibility complex (MHC) I [245,246]. SNPs affecting ERAP1 and ERAP2 function have been previously associated with various immune disorders including psoriasis [247], Behcet's disease [248]. Franke *et al.* [249] reported a *cis-eQTL* of IBD risk SNP rs2549782 with *ERAP2,* which is in high LD (r^2 = 0.76) with eQTLs SNP rs10044354 reported in our results [249]. Furthermore, Jostin's *et al.* prioritised *ERAP1* and *ERAP2* as genes potentially involved in IBD pathogenesis [92]. *GSDMB* (Gasdermin B) is a part of the gasdermin-family, which have been implicated to regulate the gastric epithelial cell apoptosis cascade through TGF-β signalling [250]. Cells transduced to express GSDM showed to induce apoptosis compared to un-transduced cells in a colony formation experiment [250]. Although the exact function of GSDMB is unknown, a recent study reported that a polymorphism in *GSDMB* is linked to an increased risk of asthma and IBD [251]. *FADS2* (Fatty acid desaturase 2) encodes an enzyme involved in the conversion of linoleic acid to pro-inflammatory arachidonic acid [252]. Franke *et al.* [249] and Peters *et al.* [253] both reported a *cis*-eQTL for *FADS2* with rs102275, which is in high LD (r^2 = 0.93) with rs174535 identified in our study. The eQTLs show an increase expression of *FADS2* within individuals with the minor allele [249,253], which is contradictory to observation that *FADS2* knockdown mice develop duodenal and ileocecal ulcerations [254].

Furthermore, 5 *cis*-eQTL, are novel discoveries within colonic tissue samples although they have been previously reported to be present in other tissues [92,127]. The 5 *cis*-eQTL were associated with genes *C2orf74* (Chromosome 2 Open Reading Frame 74), *NIPSNAP1* (Nipsnap Homolog 1), *CEP192* (Centrosomal Protein 192), *PROCR* (Protein C Receptor) and *GALC* (Galactosylceramidase) were identified within tissues including lymphocytes, fibroblasts, spleen, thyroid, adipose tissue, oesophagus epithelial cells, and artery.

6.7.2 Novel *cis-e*QTLs

By utilising GTEx, it was established that 28 (23+5) here identified *cis*-eQTLs were previously reported within colonic or other human tissues. The remaining four out of 32 *cis*-eQTLs include associations with changes in gene expression of *BORCS7* (Bloc-1 Related Complex Subunit 7), *MAP4K2* (Mitogen-Activated Protein Kinase Kinase Kinase Kinase 2), *UQCR11* (Ubiquinol-Cytochrome C Reductase, Complex III Subunit XI) and *IGLVI-70* (Immunoglobulin Lambda Variable (I)-70). These have not previously been reported in any tissue and therefore add to the novel discoveries of this study. Most notably, *MAP4K2* and *IGLVI-70* can functionally be linked to processes known to be important in IBD manifestations. MAP4K2 is a serine/threonine-protein kinase and an essential component of the MAP kinase signal transduction pathway. MAP4K2, activated by TNF-α and pro-inflammatory stimuli, is an upstream activator of stress-activated protein kinase/c-Jun N terminal kinase (JNK) signalling pathway. JNK activation has been shown to be important in intestinal inflammation in IBD [243]. IGLVI-70 is a variable of the lambda chain of immunoglobulin molecule. Immunoglobulins, antibodies, are produced by B-cells and involved in foreign antigen recognition and innate immune responses such as phagocytosis and the complement system, processes highly relevant to IBD and a healthy immune response [244]. Whereas, BORCS7 is a subunit of protein complex BORC which regulates intracellular lysosomes trafficking and positioning, interference with BORC results in collapse of lysosomes and

reduction in cell spreading and mobility [255] and UQCR11 is part of a protein complex involved in the mitochondrial respiratory chain. UQCR11 may function as an iron-sulfur protein binding factor. Although, their functional influence on IBD might not immediately be obvious but the presence of significant IBD risk SNP associated *cis*-eQTLs within colonic tissue warrants further investigation of the genes.

Furthermore, here 15 significant colonic *cis*-eQTL genes were identified within 12 IBD loci where previously published eQTL and functional data repositories had failed to help identify potential candidate genes [92,107,108,112]. Most notably, *SFMBT1* a gene associated with a lead IBD variant rs9847710 at chr3: 52978418-53142980. SFMBT1 is a histone binding protein which mediates the recruitment of corepressor complexes to target genes involved in myogenesis and antigen recognition. Although, IBD locus 3.04 did not contain any previously prioritised genes as per the most recent studies by the IBD consortium, Sing *et al.* [132] also reported the eQTL within *SFMBT1* in ileal tissue in 2015. Overall, they identified 11 *cis*-eQTLs associated with IBD SNPs, of which 5 in rectal and 6 in ileal tissue of 39 cases and 33 controls [132]. In our generated colonic data, 4 of Sing *et al.* reported eQTLs were replicated. The inability to identify 7 out 11 *cis*-eQTLs considering this study has a larger sample size, could be contributed to limited coverage of SNPs across the IBD loci within our genotype data or the observed eQTLs might be rectal or ileal specific.

7. Deconvolution of intestinal biopsy composition

7.1 Tissue heterogeneity in sequencing

While whole RNA sequencing is known to be a hypothesis-free and in-depth method to quantify gene expression, sample heterogeneity is often a concern. Expression signals measured within heterogeneous tissues are confounded by relative proportions of the cell types involved, making it challenging to determine whether variability in gene expression stemmed from differences in phenotype or tissue composition. The intestinal biopsies used to generate transcriptional data (see Material and Methods section 2.2.2) consist of a heterogeneous tissue including epithelial, stromal and various immune cell types. In order to distinguish cell-type-specific transcriptional signals and the effect of variation in cell composition on gene expression, a method for deconvolution of biopsy composition utilising gene expression data was employed.

7.2 Cellular phenotyping of biopsies

In Chapter 3.5 it was shown that high quality whole RNA sequencing data was generated from intestinal biopsies. In addition to generating transcriptomics data, 3-4 biopsies per patient sample were used to assess cellular composition. In order to assess the level of heterogeneity within the intestinal biopsies, cell populations were phenotyped by flow cytometry (see Material and Methods section 2.2.8). Epithelial cell numbers as well as various subsets of leukocytes were assessed, with epithelial cells hypothesised to make up the largest fraction of the intestinal biopsies, and leukocytes known to play an important role in inflammatory responses in the gut and IBD.

7.2.1 Gating strategy

The presence of auto-fluorescence and unspecific staining was observed in the samples. Considering the biopsy samples went through manual and enzymatic separation prior to antibody staining, this was not unexpected. The gating strategy was adapted to exclude these unspecific signals (**Figure 7.1**).

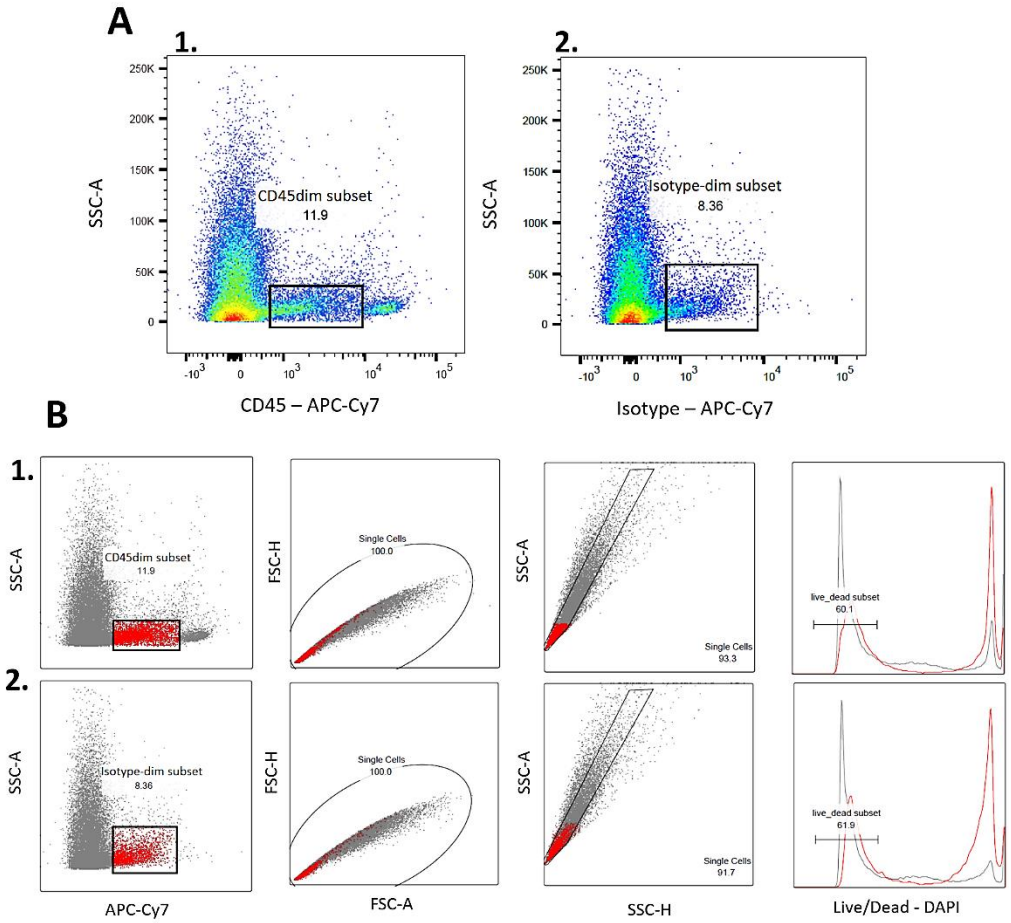


Figure 7.1 | Backgating to identify auto-florescence
CD45-APC-Cy7 (**A.1**) and Isotype control-APC-Cy7 (**A.2**) satin, with gating on a dim positive population. Backgating indicated where dim labelled population are located within previous gates (red cells) within total population of cells (grey) for CD45-APC-Cy7 (**B.1**) and Isotype control-APC-Cy7 stain (**B.2**).

Through the use of isotype controls, an APC-Cy7^{dim} population was identified which showed consistent auto-fluorescence and unspecific staining (**Figure 7.1A.2**). Backgating indicated this population contained small, low density cells with varying levels of DAPI uptake (**Figure 7.1B**). The observed APC-Cy7^{dim} population showed relative clean borders and thus the gating strategy was adjusted to exclude this subpopulation of cells from further analysis.

A gating strategy was optimised to allow for the accurate identification of proportionate numbers of epithelial cells (CD326^{pos}), leukocytes (CD45^{pos}), T helper cells (CD4^{pos}), cytotoxic T cells (CD8^{pos}), monocytes (CD14^{pos}), macrophages (CD68^{pos}) and neutrophils (CD66b^{pos}) within in the large intestinal biopsies (**Figure 7.2**).

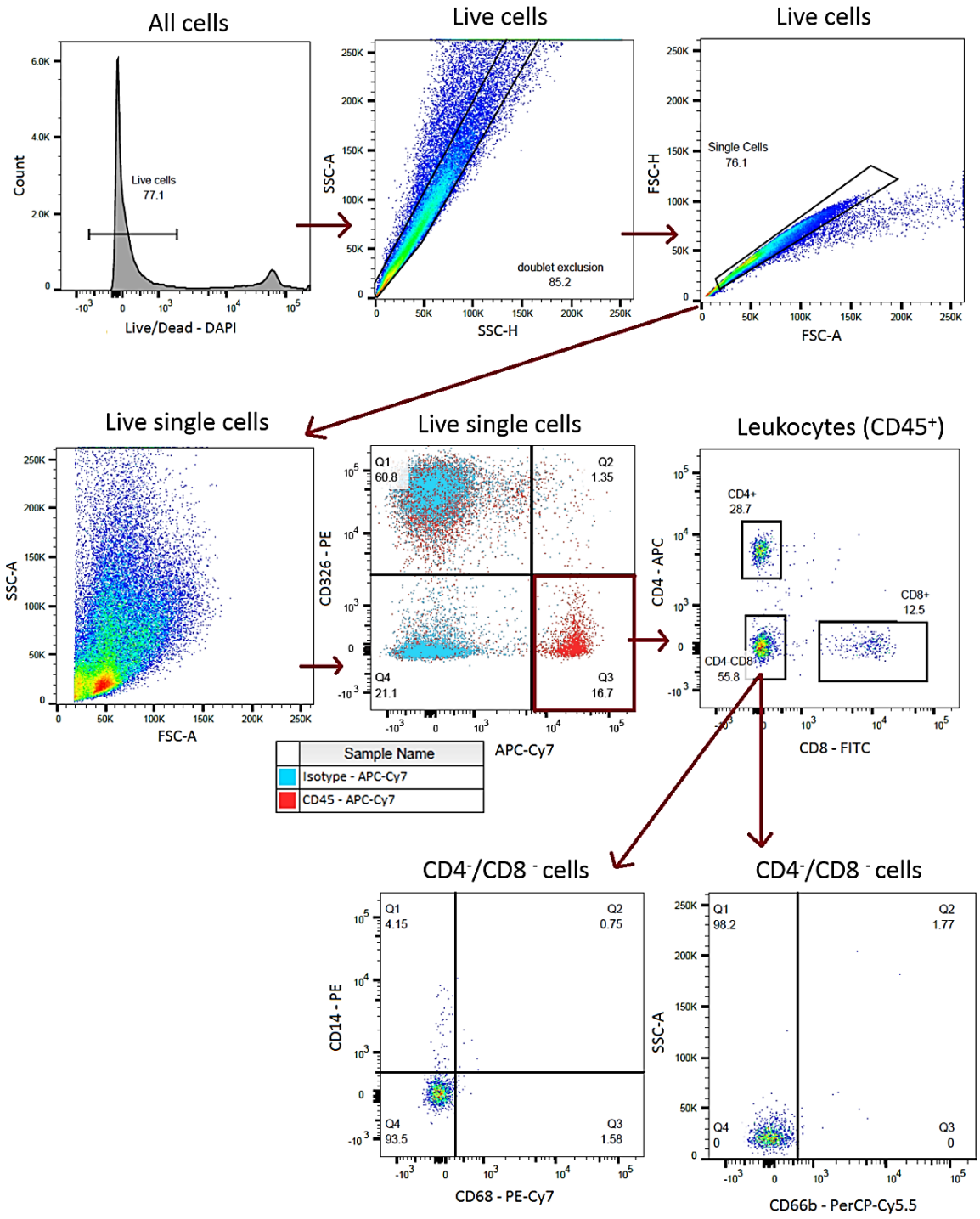


Figure 7.2 | Gating strategy cellular phenotyping biopsies
Gating strategy to identify cellular subpopulations within the large intestinal biopsies. Live cells were identified on basis of DAPI exclusion and cell doublets were gated out using SSC-A vs SSC-H and FSC-H vs FSC-A. Epithelial cells were $CD45^{neg}/CD326^{pos}$ and Leukocytes were $CD45^{pos}/CD326^{neg}$ with the positive stain based on isotype control staining. Identified leukocytes were further subdivided identifying $CD4^{pos}$ T helper cells and $CD8^{neg}$ cytotoxic T cells. $CD45^{pos}/CD4^{neg}/CD8^{neg}$ subpopulation identified Monocytes by $CD14^{pos}$, Macrophages $CD68^{pos}$ and neutrophils $CD66b^{pos}$

An isotype control for APC-Cy7 was included in an FMO (Fluorescence minus one) antibody cocktail, to enable accurate gating on the CD45 positive leukocyte cell population. Leukocyte subpopulations $CD4^{pos}$, $CD8^{pos}$ and $CD4^{neg}/CD8^{neg}$ showed clear borders when gating and thus no further isotype controls were required (**Figure 7.2**). $CD14^{pos}$ monocytes were observed to make up only a minor proportion of leukocytes present in uninflamed intestinal biopsies (**Figure 7.2**). It is possible that influx of $CD14^{pos}$ monocytes occurs only during an active inflammatory response in the gut. In response to the observed low numbers of $CD14^{pos}$ monocytes within the biopsy samples, a CD68 macrophage stain was introduced to assess the presence of macrophages that had lost their $CD14^{pos}$ expression within the gut. Low abundance of CD68 macrophages was observed (**Figure 7.2**). Furthermore, CD66b, a neutrophil marker, was introduced to assess the level of inflammation within a subset of the intestinal biopsies (n=4). An uninflamed state of the biopsy tissue was observed through the absence of a positive stain for CD66b (**Figure 7.2**).

7.2.2 Biopsy composition

The cellular composition was assessed for n=24 CD patient biopsy samples for which transcriptomics data was also generated. The most abundant cell subtypes were epithelial cells ($CD45^{neg}/CD326^{pos}$) with median abundance of 60% of live single cells and leukocytes ($CD45^{pos}$) contributing 14% of total live cells (**Figure 7.3A**). $CD4^{pos}$ T helper cells were observed to be the most abundant cell-type (median 21%) within the leukocyte subset, with $CD8^{pos}$ cell (median 16.2%) being a close second. A very low abundance of $CD14^{pos}$ monocytes, at approximately 0.5%, was observed (**Figure 7.3B**).

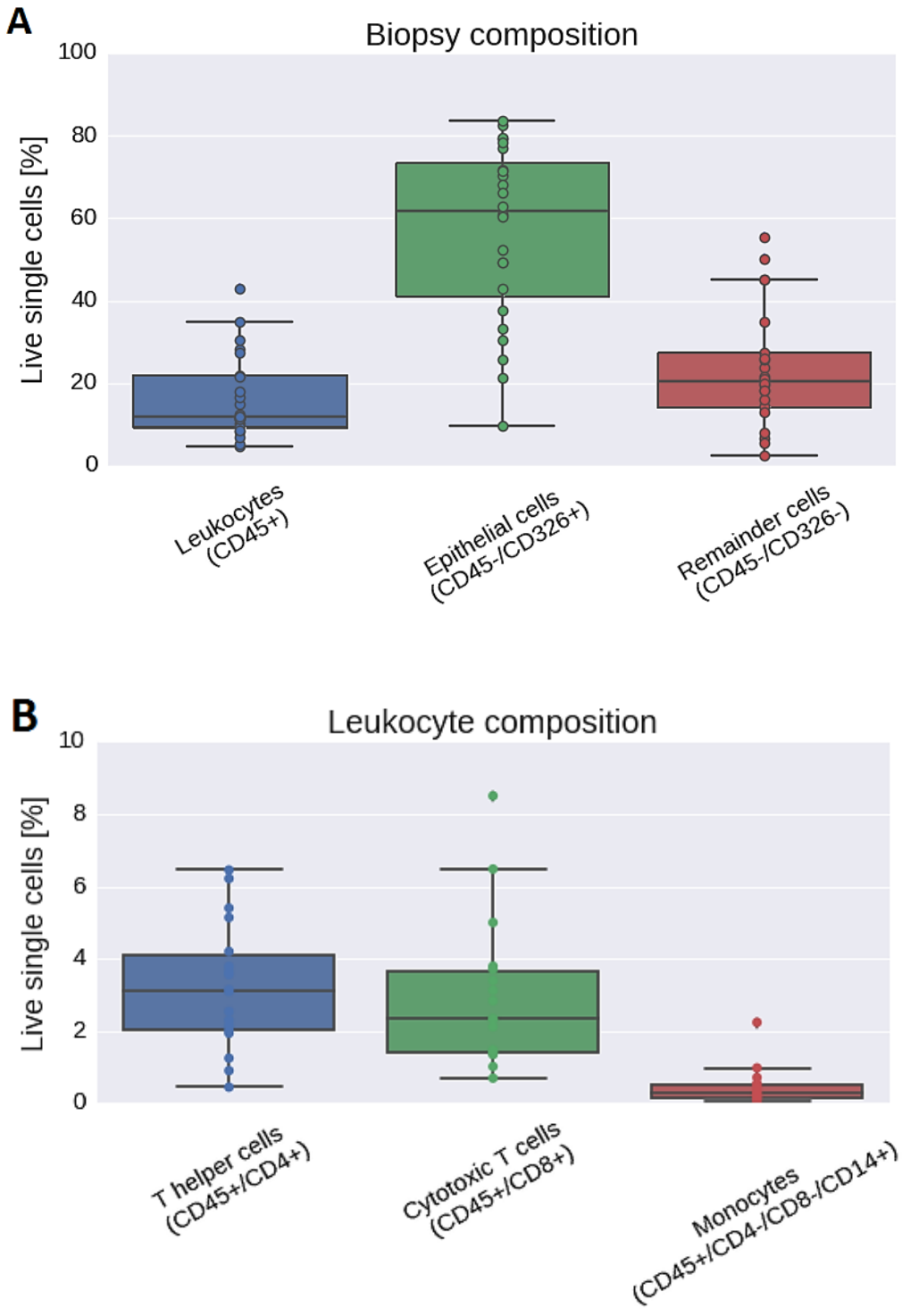


Figure 7.3 | Cellular phenotype of intestinal tissue biopsies by FACS
Box and whiskers plot visualising collective flow cytometry data plotted as % positive live cells for epithelial and leukocyte cell subsets (**A**) and leukocyte immune cell subtype (**B**). The box representing the 25th and 75th percentile from the median and the whiskers repressing the lowest and highest value. n=24.

Substantial variation within cellular composition of the biopsy samples was observed, with epithelial cell proportions varying from 10% to 82% and the leukocyte population accounting for 5% to 42% of live cells (**Figure 7.3A**). This, highlights the importance of investigating composition of heterogeneous samples used for transcriptional analyses. Variation could be due to inter-patient variation or the location or depth at which the biopsy was taken. Less variation was observed within the leukocyte subpopulations, although $CD4^{pos}$ and $CD8^{pos}$ percentages varied from 1% to 7% and 9%, respectively (**Figure 7.3B**).

7.3 Deconvolution of biopsy composition

In 3 out of 24 biopsy samples which were phenotyped for cellular composition using FACS, no gene expression data was generated owing to low quality RNA (average RIN = 2.3). A total of 15,517 transcripts were detected above background within the colonic intestinal biopsies (see Chapter 4.1) of the 21 CD patients for whom cellular composition was assessed. A univariate analysis was employed to identify the genes which had a significant influence on biopsy composition. The genes identified as significant were advanced into a penalised regression, identifying a set of genes which collectively predict the cell count for each cell type. This 'predictive gene set' was employed to predict the percentage of each of the known phenotyped cell types within the n=21 biopsy subset, enabling us to correlate the predicted values with the known values. Finally, the 'predictive gene set' was used to deconvolute the cellular composition of the remaining biopsies. The below analyses was designed and executed by Seth Seegobin, a PhD student within the statistical genetics unit.

7.3.1 Univariate analysis using a marginal model

To account for the fact that relative proportions of cell types are correlated; when one goes up another has to come down, covariates between the cell types were calculated into an unstructured covariate matrix using a marginal model. The unstructured covariate matrix together with the cell type, cell frequencies and normalised gene expression count values were fitted into a univariate

analysis using SAS (statistical analysis software). This generated a p-value per gene, indicating the gene's influence on cell type composition of the biopsy per patient. Out of the 15,517 genes expressed, 1,725 genes (1.11%) were identified to significantly ($p<3.2x10^{-6}$) contribute to biopsy composition (**Figure 7.4**).

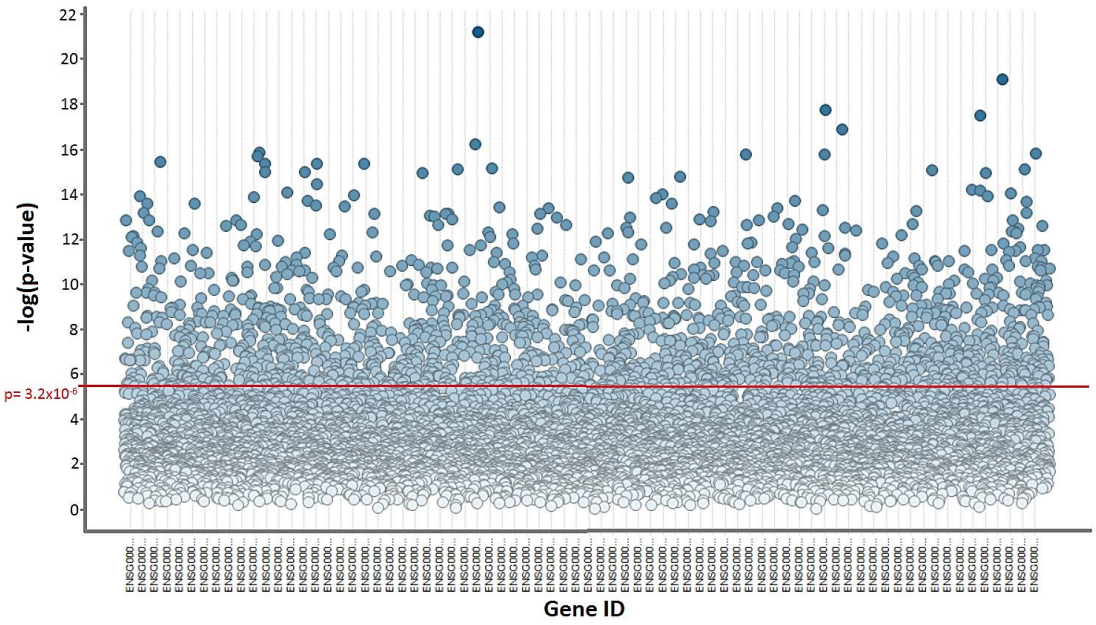


Figure 7.4 | Contribution of gene expression on cell type
Gene ID (x-axis) versus influence of expression levels on intestinal biopsy cell composition (y-axis). With each dot representing a transcript (n=15,517) and a significance cut-off line at p= $3.2x10^{-6}$ (red line).

7.3.2 Machine learning penalised regression

The 1.11% of genes significantly associated with intestinal biopsy composition were identified, and progressed into a multivariate analysis to identify which set of significant genes could collectively predict cell count per cell type, using lasso (least absolute shrinkage and selection operator). The cell types predicted by gene expression included epithelial cells ($CD45^{neg}/CD326^{pos}$), leukocytes ($CD45^{pos}$), T helper cells ($CD45^{pos}/CD4^{pos}$), cytotoxic T cells ($CD45^{pos}/CD8^{pos}$) and monocytes ($CD45^{pos}/CD14^{pos}$). Lasso identified 20 genes (**Table 7.1**) which could collectively predict the percentage of each above mentioned cell type contributing to biopsy composition.

Table 7.1 | Predictive genes

Ensemble ID	Gene Name	Function
ENSG00000011009	*LYPLA2*	lysophospholipase II, regulates multifunctional lysophospholipids within biological membranes
ENSG00000011083	*SLC6A7*	Solute carrier family 6 (neurotransmitter transporter), member 7, functions as a L-proline transporter protein in the brain
ENSG00000011275	*RNF216*	ring finger protein 216, inhibits TNF- and IL-1 induced NF_kB activation.
ENSG00000012779	*ALOX5*	arachidonate 5-lipoxygenase, plays a role in leukotrienes synthesis, which mediate a number of inflammatory and allergic conditions.
ENSG00000012822	*CALCOCO1*	a coactivator for aryl hydrocarbon and nuclear receptors, involved in cellular metabolism, protein synthesis and degradation.
ENSG00000013275	*PSMC4*	Involved in 26S proteasome assembly, which enables ATP-dependant degradation of ubiquitinated proteins
ENSG00000013364	*MVP*	major vault protein, involved in signal transduction and nucleo-cytoplasmic transport
ENSG00000013441	*CLK1*	CDC-like kinase 1, involved in pre-mRNA processing
ENSG00000013503	*POLR3B*	DNA directed RNA polymerase III
ENSG00000013563	*DNASE1L1*	Protein part of the deoxyribonuclease family
ENSG00000014216	*CAPN1*	Calcium-regulated non-lysosmal protease which catalyses substrates involved in signal transduction
ENSG00000014257	*ACPP*	Tyrosine phosphatase that dephosphorylates various substrate under acidic conditions
ENSG00000015133	*CCDC88C*	coiled-coil domain containing protein, negatively regulator of Wnt signalling pathway
ENSG00000019505	*SYT13*	Synaptotagmin, may be a transport vesicle involved in calcium ion binding and calcium-dependent phospholipid binding
ENSG00000021355	*SERPINB1*	Intercellular inhibitor of granzyme H, protecting tissue for damage at inflammatory sites.
ENSG00000023191	*RNH1*	ribonuclease/angiogenin inhibitor
ENSG00000023445	*BIRC3*	Protein regulated caspases and apoptosis, modulates inflammatory signals and immunity
ENSG00000023516	*AKAP11*	A-kinase anchor protein, binds to regulatory subunits of protein kinase A and confines these to locations within the cell
ENSG00000023608	*SNAPC1*	Part of the SNAPc complex, required for transcription of RNA polymerase II and III
ENSG00000025708	*TYMP*	Promotes angiogenesis and stimulated growth of endothelial cells

For each of the cell types, a set of 16 – 20 genes and the intercept were identified and an estimate score was generated (**See appendix 6**). The predicted fractions of each cell type within an intestinal biopsy sample can be calculated by taking the sum of the estimate score multiplied by the gene count value for all genes associated with the cell type plus the intercept estimate value (see Material and Methods section 2.2.10.3). Cell type fractions were estimated for all 21 samples used to build the prediction model allowing comparison between FACS observed cell type percentages and predicted cell type fractions (**Table 7.2**).

Table 7.2 | Observed and predicted percentages per cell type

Sample ID	Observed CD326+	Predicted CD326+	Observed CD45+	Predicted CD45+	Observed CD4+	Predicted CD4+
GKT1908	49.3%	0.493	21.8%	0.218	4.2%	0.042
GKT2255	30.5%	0.305	15.0%	0.15		
GKT1914	42.9%	0.429	10.3%	0.103	2.1%	0.021
GKT2878	71.8%	0.718	12.4%	0.124	1.9%	0.019
GKT2879	70.3%	0.703	4.8%	0.048	0.9%	0.009
GKT2697	37.7%	0.377	5.3%	0.053		
GKT2168	82.6%	0.826	8.4%	0.084		
GKT2922	76.9%	0.769	11.4%	0.114	0.5%	0.0045
GKT2059	25.8%	0.258	35.0%	0.35		
GKT1525	21.4%	0.214	28.2%	0.282	5.4%	0.054
GKT2329	49.2%	0.492	21.5%	0.215	3.6%	0.036
GKT3089	68.0%	0.68	12.0%	0.12	2.5%	0.025
GKT2689	79.6%	0.796	6.9%	0.069	1.3%	0.0125
GKT3084	52.3%	0.523	27.3%	0.273	6.5%	0.0645
GKT0327	60.8%	0.608	16.7%	0.167	3.1%	0.031
GKT0711	60.5%	0.605	11.2%	0.112	3.6%	0.0356
GKT2790	33.1%	0.331	30.5%	0.305	5.1%	0.0513
GKT2190	83.7%	0.837	9.4%	0.094	3.8%	0.0378
GKT3080	66.1%	0.661	11.2%	0.112	2.3%	0.02299
GKT2126	62.9%	0.629	18.0%	0.18	6.2%	0.0621
GKT3081	78.5%	0.785	12.0%	0.12	2.0%	0.0201

Sample ID	Observed CD8+	Predicted CD8+	Observed CD14+	Predicted CD14+
GKT1908	1.0%	0.01019	0.5%	0.0042
GKT2255				
GKT1914	1.5%	0.0148	2.2%	0.022
GKT2878	6.5%	0.0648	0.2%	0.00132
GKT2879	2.8%	0.0284	0.1%	0.00188
GKT2697				
GKT2168				
GKT2922	5.0%	0.05	0.1%	0.00049
GKT2059				
GKT1525	1.4%	0.0141	0.2%	0.0021
GKT2329	2.3%	0.0234	0.1%	0.0012
GKT3089	1.4%	0.0141	0.3%	0.0037
GKT2689	2.1%	0.0212	0.2%	0.0015
GKT3084	8.5%	0.0850	0.3%	0.0035
GKT0327	1.4%	0.0135	0.3%	0.0029
GKT0711	1.4%	0.0142	0.3%	0.0027
GKT2790	3.7%	0.0370	0.5%	0.0054
GKT2190	2.3%	0.0233	0.4%	0.0044
GKT3080	3.8%	0.0379	1.0%	0.0096
GKT2126	3.1%	0.0312	0.2%	0.0018
GKT3081	3.4%	0.0339	0.7%	0.0062

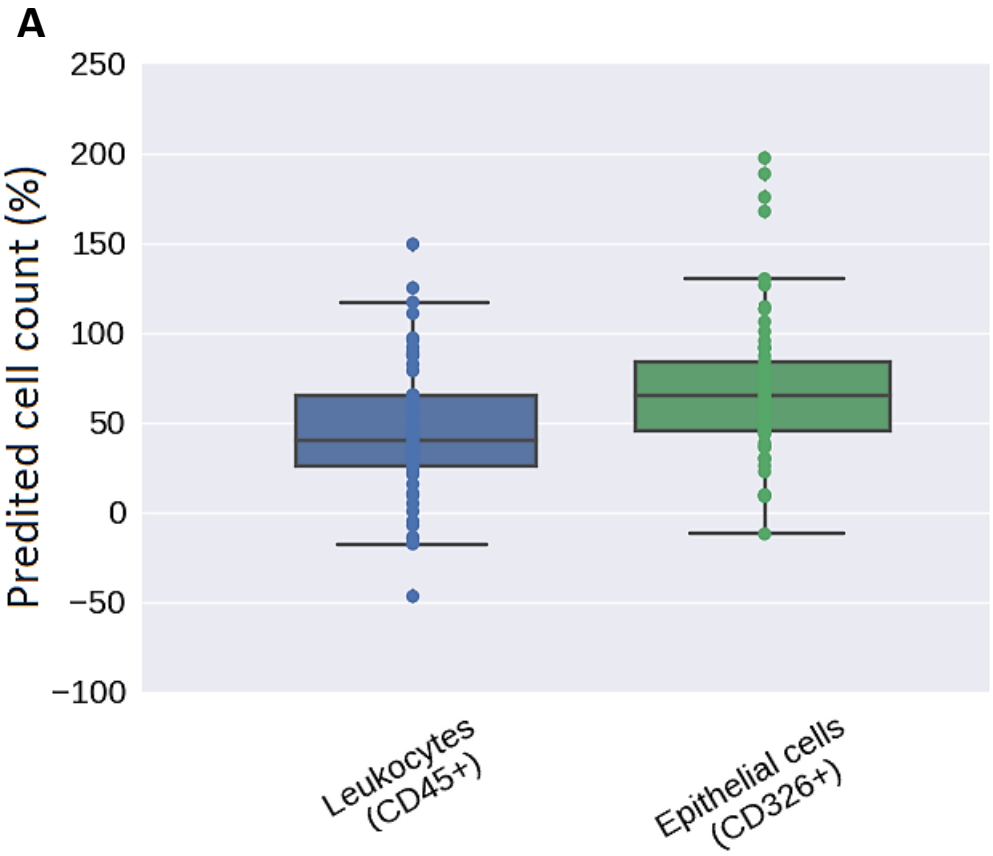
* empty cells indicate missing data

The predictive model achieved a 100% prediction between the observed and predicted cell type fractions for all cell types with exception of $CD14^{pos}$ monocytes (**Table 7.2**). The $CD14^{pos}$ monocyte population exhibited 78% prediction between the observed and predicted percentage, with GKT2922 predicted 0.049% and observed 0.1% being the largest deviation (**Table 7.2**). It was hypothesised that this was due to the low abundance of $CD14^{pos}$ monocytes (less than 1% to the overall biopsy composition). Cellular phenotype data for the leukocyte subsets $CD4^{pos}$, $CD8^{pos}$ and $CD14^{pos}$ was not collected within the early processed samples.

The 100% predictive power of the model demonstrates that gene expression data can be employed for the deconvolution of heterogeneous tissues.

7.4 Deconvolution of biopsy composition

In the above section the proportions of epithelial cells ($CD326^{pos}$), leukocytes ($CD45^{pos}$), T helper cells ($CD4^{pos}$), cytotoxic T cells ($CD8^{pos}$) and monocytes ($CD14^{pos}$) within intestinal tissue biopsies were predicted with a 100% accuracy, based on gene expression levels of 20 key genes. By applying this method to the 57 CD intestinal biopsy samples, for which gene expression data was generated, the aim was to deconvolute the composition of these RNA sequenced biopsy samples (**Figure 7.5**).



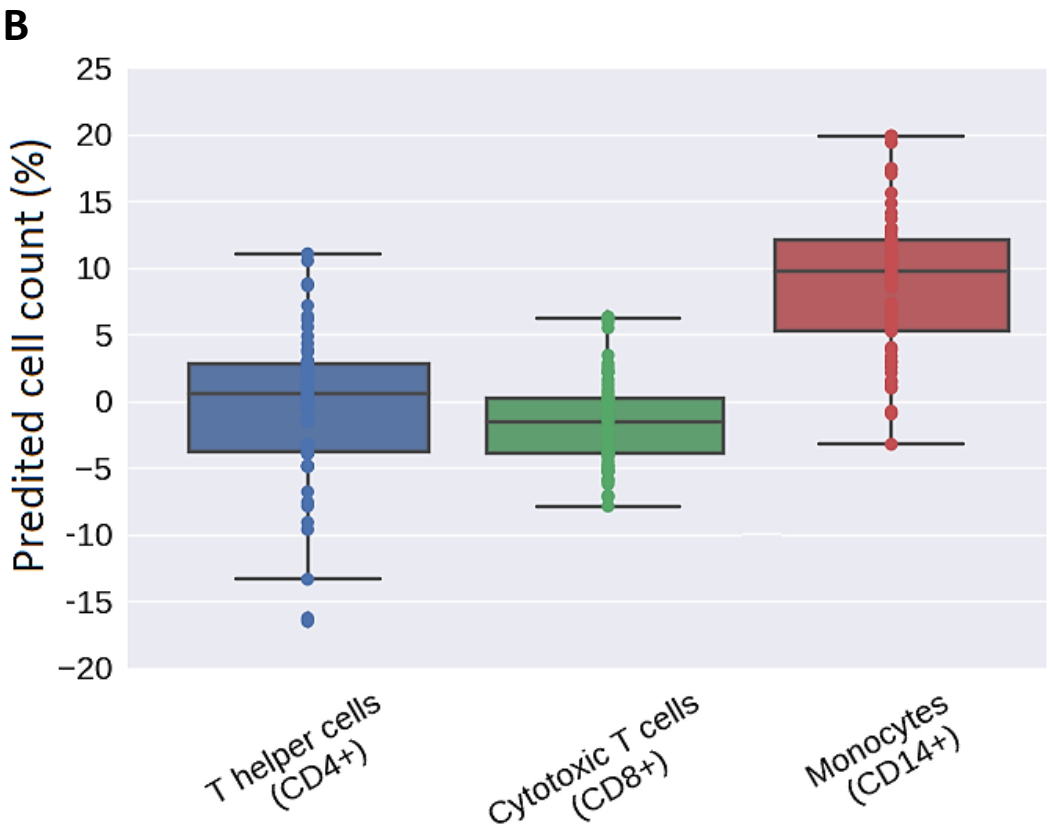
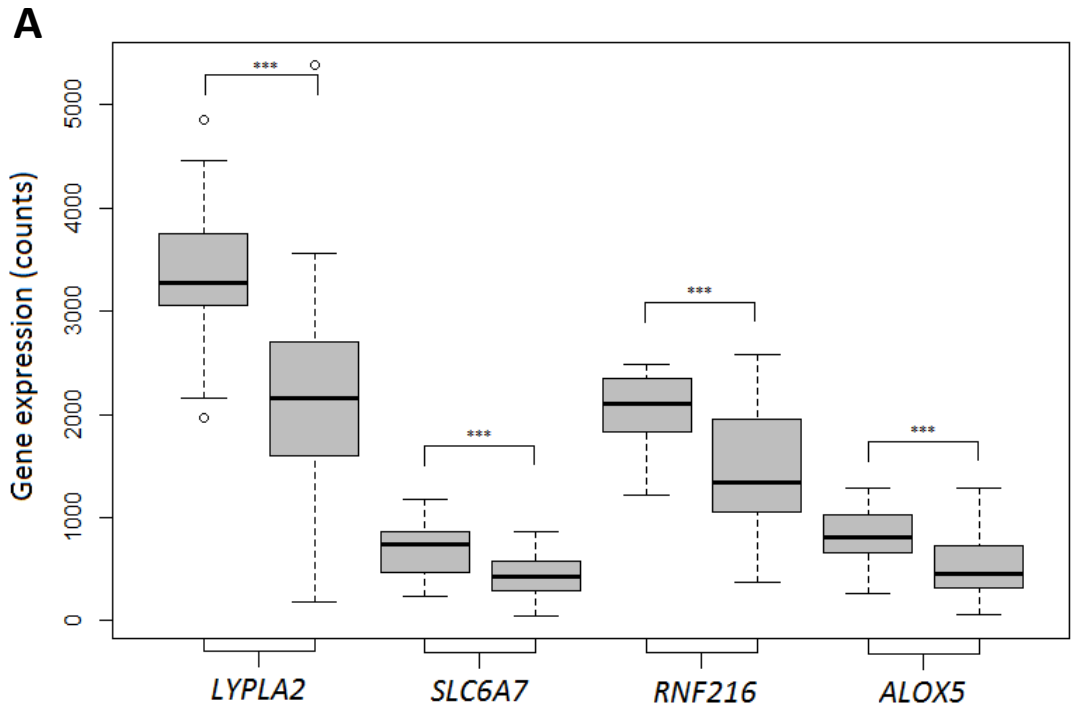


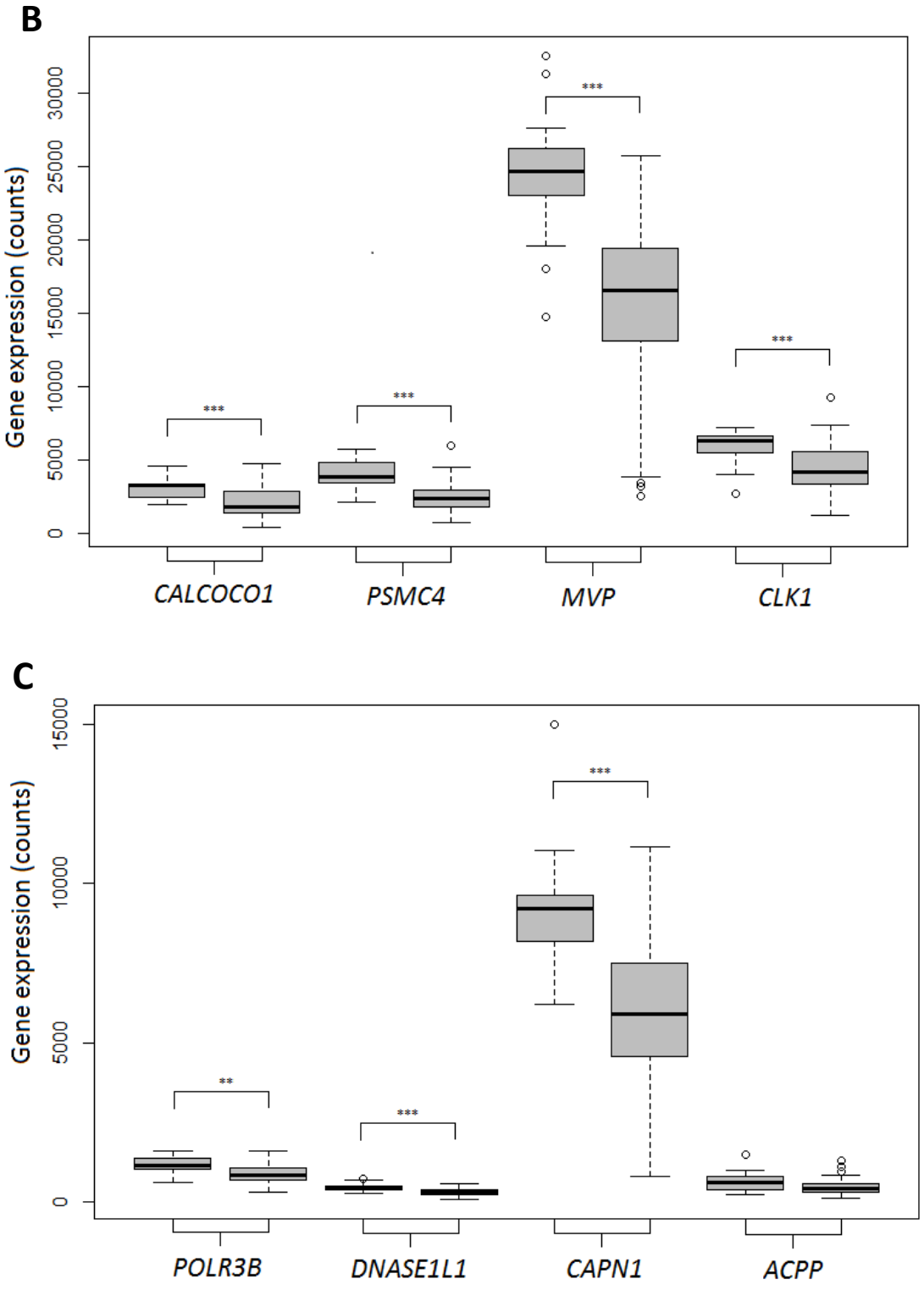
Figure 7. 5 | Cell type predictions based on gene expression
Box and whiskers plot visualising the predicted percentage of CD45^{pos} leukocytes and CD326^{pos} epithelial cells (**A**) and leukocyte cell subtypes CD4^{pos}, CD8^{pos} and CD14^{pos} (**B**) within 57 CD biopsies based on their gene expression. The box representing the 25th and 75th percentile from the median and the whiskers repressing the lowest and highest value.

The upper and lower quartiles of the predicted fractions of CD45^{pos} leukocytes and CD326^{pos} epithelial cells were observed to fall within a biologically realistic range (0% -100%), with the whiskers - indicating outliers - showing < 0% or >100% values (**Figure 7.5A**). The median CD45^{pos} leukocyte estimated fraction was predicted to be 39.7% (**Figure 7.5A**) which was higher than the 14% observed median CD45^{pos} leukocyte cell population within the 24 samples that underwent cell phenotyping (**Figure 7.3A**). The median predicted value for the CD326^{pos} epithelial cell fraction was estimated to be 65% (**Figure 7.5A**), which is very similar to the median of 61% observed within the phenotyped set (**Figure 7.3A**). Predictions for the leukocyte immune subtypes appear more challenging, which could be due to their small overall contribution to biopsy composition. CD14^{pos} monocytes, with median 9.7%, were predicted to be the

most frequent immune-cell subtype (**Figure 7.5B**), whereas within the phenotyped samples their observed frequency was much lower, with a median of 0.3% (**Figure 7.3B**). $CD4^{pos}$ cells were predicted to be more abundant than $CD8^{pos}$ cells (**Figure 7.5B**), which is consistent with the phenotype data (**Figure 7.3B**). However, $CD8^{pos}$ cells were predicted to exhibit a median of -1.5% which is biologically impossible (**Figure 7.5B**). Although, the model returns feasible predictions for the more abundant cell populations: $CD45^{pos}$ leukocytes and $CD326^{pos}$ epithelial cells, biologically-impossible (<0% or >100%) predictions are generated within the low abundant immune cell types. This indicates that the prediction model is not as accurate as suggested by the predictions of the 21 cellular phenotyped samples (**section 7.3.2**).

To investigate a potential cause for the apparent reduced accuracy of the predictive model within the tested 57 CD biopsy samples, mean expression values of the 20 predictive genes were compared with the 21 phenotyped samples (**Figure 7.6A-E**).




B
Gene expression (counts)
0
5000
10000
15000
20000
25000
30000
***
***
***
***
CALCOCO1
PSMC4
MVP
CLK1
C
Gene expression (counts)
0
5000
10000
15000
**
***
***
POLR3B
DNASE1L1
CAPN1
ACPP

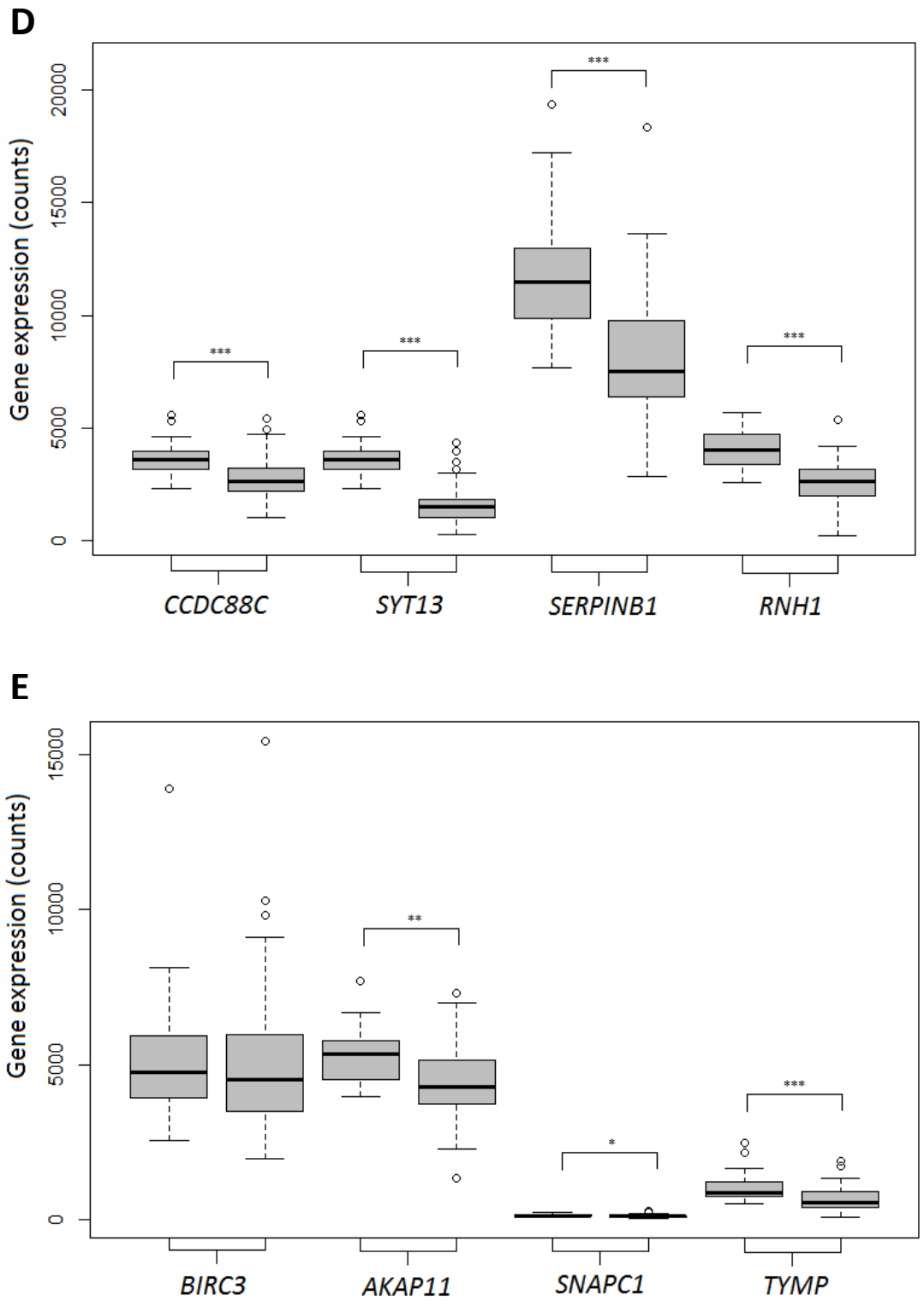


Figure 7.6 | Expression of 20 genes utilized to predict cell composition
Box and whiskers plots visualising the gene expression normalised count values of the 20 predictive genes (5 genes per plot, **A-E**) for both the 21 cellular phenotyped (left box per gene) and the 57 CD samples used in the prediction (right box per gene). The box representing the 25th and 75th percentile from the median and the whiskers repressing the lowest and highest value. Significant difference median gene expression within each gene was calculated using a two-tailed Mann-Witney test ($p<0.05$*, $p<0.005$**, $p<0.0005$***).

With the exception of *BIRC3* and *ACPP* all of the predictive genes exhibited significant difference in mean expression between the 21 samples used to build the predictive model and the 57 CD samples (**Figure 7.6A-E).** The fact that 18 out of 20 predictive genes show significant differences in expression between the two sample groups, could be contributing to the flaws in the accuracy of the cell type predictions for the 57 CD biopsies. It suggests that the 21 samples used to build the prediction model were not representative of the overall biopsy samples. The small sample size of n= 21, sequencing batch effect or the site within the colon (e.g. transverse or descending colon) could also be a factor in this.

7.5 Discussion

Whole RNA sequencing offers a high throughput, high quality way to quantify gene expression, although sample heterogeneity is often a concern. Measured expression signals within heterogeneous tissues are confounded by relative proportions of the cell types involved. This can lead to cell type specific signals being cancelled out or differences in expression being caused by sample cell composition. In order to assess sample composition associated expression differences and potentially identify cell type specific signals it is essential to address sample heterogeneity in RNA sequencing experiments. The intestinal biopsies used to generate transcriptional data (see Material and Methods section 2.2.2) consist of a heterogeneous tissue including epithelial, stromal and various immune cell types. In order to distinguish cell-type-specific transcriptional signals and the effect of variation in cell composition on gene expression, a method for deconvolution of biopsy composition utilising gene expression data was employed.

Experimental methods to resolve tissue heterogeneity have been proposed, such as laser-capture microdissection (LCM), allowing dissection of morphologically distinguishable cell types. The RNA yield and quality is often considerably lower following LCM, furthermore it can only be applied to

morphologically distinguishable tissue which in our research is not applicable. Another experimental method to address tissue heterogeneity is cell purification through bead based or flow cytometry methods. Cell sorting through flow cytometry results in high quality, high purity samples, although it might prove a challenge to generate the required yield of RNA for sequencing of low abundance cell-types. Although with new and improved amplification methods becoming available, this is less of an issue. However, cell purification methods are laborious, expensive and could trigger uncontrolled processes in the cell altering transcription. Taking all this into consideration, the use of an *in silico* approach to address tissue heterogeneity has great appeal. Various methods utilising either gene expression profiles or DNA methylation status of purified cell types have been used to deconvolute heterogeneous tissue [256-259].

One of the first studies to attempt this using gene expression profiling was D. Venet *et al.* [260]. They employed a linear model to predict cell-type proportions within colonic cancer biopsies based on 'marker genes' which are uniquely expressed in each cell type including muscle cells, fibroblasts and macrophages. The model was built upon the assumption that the expression of each gene in a heterogeneous sample is the weighted average of the expression levels existing in pure populations of those cells [260]. Computational models have evolved over the years but the majority still rely on pre-determined cell-type-specific expression profiles from a range of pure/single-cell-type found within the whole tissue [256,261,262]. Deconvolution models, focusing on peripheral blood cells, where neither proportions of cells nor signature gene expression data was available have been proposed [263]. By using known positive and negative marker genes specific to each cell type within peripheral blood their 'pure' expression could be estimated from the mixed RNA sequencing data [257,263]. This method could potentially be employed to estimate the proportion of leukocytes subsets within our intestinal data but 'marker genes' were not available for the other major cell types in intestinal mucosa, i.e. epithelial or stromal cells meaning it had limited use.

Due to the large number of different cell types known to be present in intestinal biopsies in varying quantities, we set out to find a more suitable approach to quantifying them in a large number of samples. This was done using RNAseq based gene expression data on a subset of whole biopsies from which the proportions of cells by flow cytometry were previously quantified. Although, flow cytometry is a validated technology for quantifying cell proportions based on antibody based fluorescent labelling; it should be taken into consideration that the biopsy samples underwent an enzymatic digestion prior to being incubated for 2 hours to stimulated recovery of cell surface receptors. During this process a level of cell death was observed and although all biopsies were process according to the same protocol some cell subsets might be more susceptible to cell death than other cellular subtyped in the biopsy, therefore potentially skewing the data. There are currently no other methods without limitations to perform cellular phenotyping on tissue biopsies.

A new deconvolution method was derived based on the ability of a subset of genes that had demonstrated high correlation with sub-cellular proportions in a training set. It was hypothesised that these 'predictive genes' could predict cell count per cell type using a machine learning penalised regression model. The model was able to predict the cell counts of five main cell types within the intestinal biopsy samples used to build the model (n=21) with a 100% accuracy. When employing the model to predict cell count of the five cell types within intestinal biopsies not used to build the model, biologically impossible values were returned i.e. cell counts below 0% or above a 100%. In order to address the cause for these biologically impossible results returned by our deconvolution model, mean gene expression was assessed for the 'predictive genes' within the sample set used to build the model and the samples for which their composition was predicted using the model. It was established that the mean expression of the 'predictive genes' within the samples used to build the model did not represent the overall expression within the intestinal biopsy samples accurately. The difference in mean expression can most likely be contributed to the small sample size of the sample set used to build the model. Repsilber *et al.* did show that deconvolution models lacking cell-type-

specific RNA sequencing data, require a large sample size (n=120) to generate accurate predictions [263]. Furthermore, the intestinal location where the biopsy was taken and sequencing batch effects should be considered as potential causes for the inaccurate in the deconvolution model.

The developed model showed promise in addressing the much debated issue of tissue heterogneity in the field of RNA sequencing. Further optimisation and validation will need to be performed to finalise and perfect a deconvolution model.

8. Biomarkers predictive of relapse in Crohn's disease

Crohn's disease (CD) is a chronic condition characterised by episodes of relapse requiring either medical or surgical intervention to induce remission. In many patients the disease is progressive, and the ability to predict who will relapse is extremely poor despite extensive work using clinical characteristics and biomarkers. Relapse is often subclinical initially, but even at such a stage can be associated with irreversible bowel damage. The ability to predict which patients are more likely to relapse will enable targeting of expensive drugs, restricted in their availability, to the appropriate patients as well as avoiding exposing patients to unnecessary medical therapy with potentially serious side effects. It will also allow for closer monitoring of patients at higher risk of relapse, thereby minimising the burden on patients and clinical services in cases where less frequent monitoring is possible. It was therefore decided to use genome-wide microarray to derive transcriptional profiles from unstimulated and stimulated $CD4^{pos}$ T helper cells, $CD8^{pos}$ cytotoxic T cells and $CD14^{pos}$ monocytes isolated from whole blood of patients, and compare profiles in those who do and do not relapse within 1 year. With the aim of identifying transcriptional differences that could potentially elude to effective biomarkers for the prediction of relapse in patients.

8.1 Patient samples

Blood samples for isolation of peripheral blood mononuclear cells (PBMCs) were collected from 49 patients in endoscopic or clinical remission at time of recruitment. Their progress was monitored over a 12-month period to determine who did and did not relapse, and a second blood sample was collected at time of relapse or at 12-months post recruitment. Patients within three subsets were recruited:

- 20 post-surgery patients, either resection or stoma reversal surgery.
- 20 routine gastroenterology clinic patients
- 9 patients having been withdrawn from anti-TNF treatment, Humira or Infliximab

The post-surgery patients were recruited either two weeks post their resection surgery or at the day of their stoma reversal. The routine gastroenterology clinic patients were recruited when reporting clinical remission during their appointment and the anti-TNF withdrawal patients were recruited eight week following their final Infliximab injection (or two weeks for Humira). Relapse was assessed at a 6-month follow-up colonoscopy for the post-surgery patient cohort, with a Rutgeerts score ≥ 2 classified as relapse. For the anti-TNF withdrawal and gastroenterology clinic patient cohorts, relapse was classified as a need to change the patients' medication. An overall relapse rate of 37% was observed, with the highest incidence of relapse seen in the post-surgery cohort at 45% (9 out of 20). The anti-TNF withdrawal cohort exhibited 33% relapse rates (3 out of 9) and 30% of gastroenterology clinic patients (6 out of 20) relapsed (**Table 8.1**).

Table 8.1 | Patient demographics

	Post-surgery patients	Anti-TNF withdrawal patients	Gastro-clinic patients
Mean age	32.7	33.3	44.8
# Female	14	5	8
# Male	6	4	12
# Relapsed	9	3	6
within	5 - 9 months	2-8 months	1-12 months
# Remission	11	6	14
beyond	> 6 months	>12 months	>12 months

8.2 Cell sorting

An average of $6.45x10^7$ PBMCs were extracted from 50 ml of blood stored for between 2 and 12 months. An average of $4.7x10^6$ live PBMCs were retrieved, with samples having been stored for longer periods of time showing lower viability. Thawed and rested PBMCs were stained with fluorescent labelled antibodies (see Materials and Methods section 2.3.4) and flow cytometry based cell sorting was employed to isolated $CD4^{pos}$ T helper cells, $CD8^{pos}$ cytotoxic T

cells and $CD14^{pos}$ monocytes. The purity of the separated cells was then assessed by flow cytometry (**Figure 8.1**). Prior to cell sorting a mixed PBMC composition within the samples was observed with approximately 10% monocytes ($CD14^{pos}$), 42% cytotoxic cells ($CD8^{pos}$) and 19% T helper cells ($CD4^{pos}$) of live leukocytes (**Figure 8.1A**). Post cell sorting cell population purities of 99.6% for $CD4^{pos}$, 99.4% for $CD8^{pos}$ and 96.5% for $CD14^{pos}$ cells was observed **(Figure 8.1B-D**).

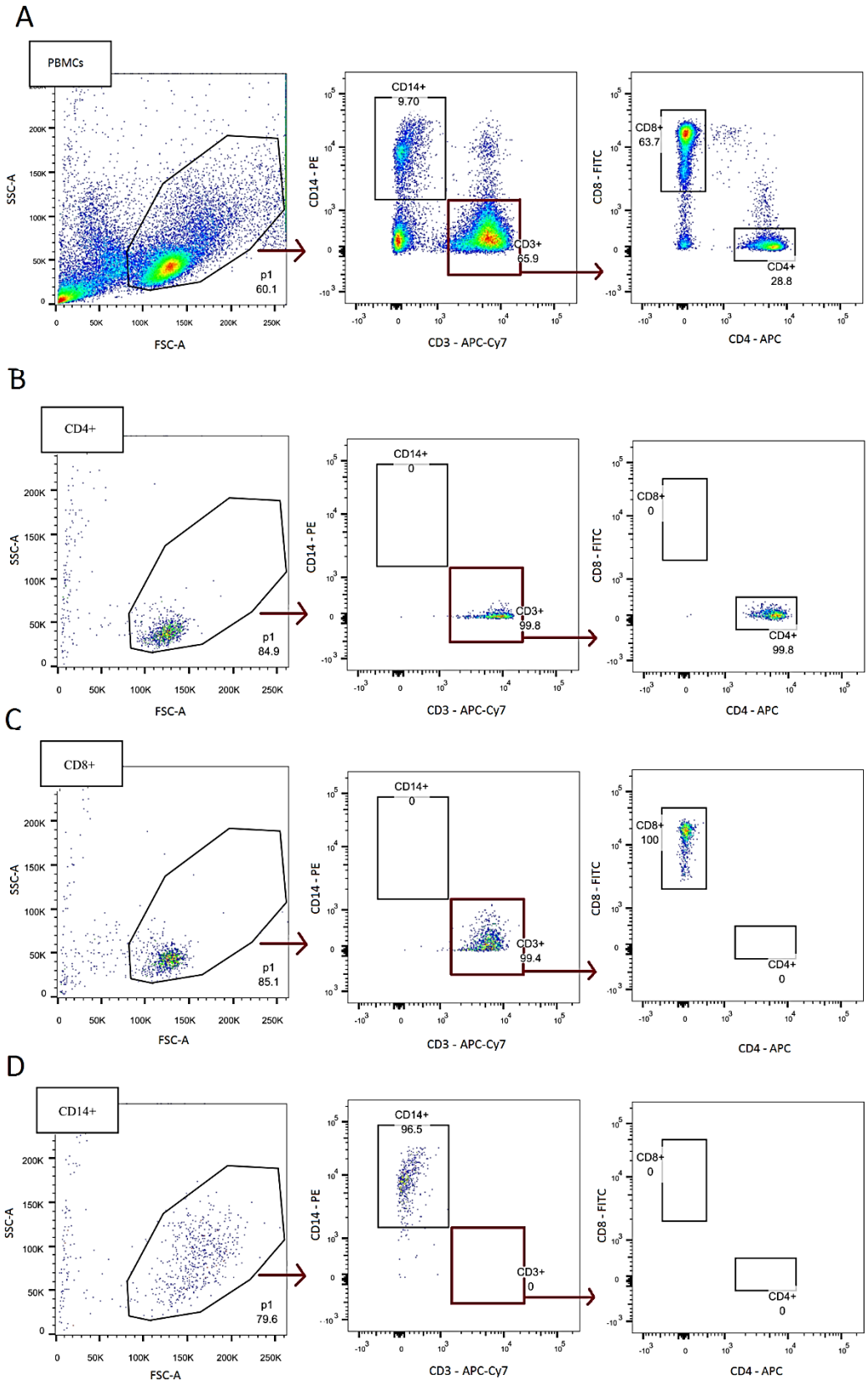


Figure 8.1 | Flow cytometry based purity check
Gating strategy to identify cellular subpopulations within the peripheral blood mononuclear cells (PBMCs) pre and post flow cytometry based cell sorting. Leukocytes were identified using SSC-A vs FSC-A gating. Monocytes were identified as $CD14^{pos}$ (**B**) and T helper and cytotoxic T cells were $CD3^{pos}$ and $CD4^{pos}$ (**C**) or $CD8^{pos}$ (**D**), respectively.

Average purities achieved for the $CD4^{pos}$ and $CD8^{pos}$ cell populations within the 55 processed samples reached 99.5% (**Figure 8.2**). $CD14^{pos}$ monocytes showed the highest variation in purity, although still reaching an average of 97.7% purity (**Figure 8.2**).

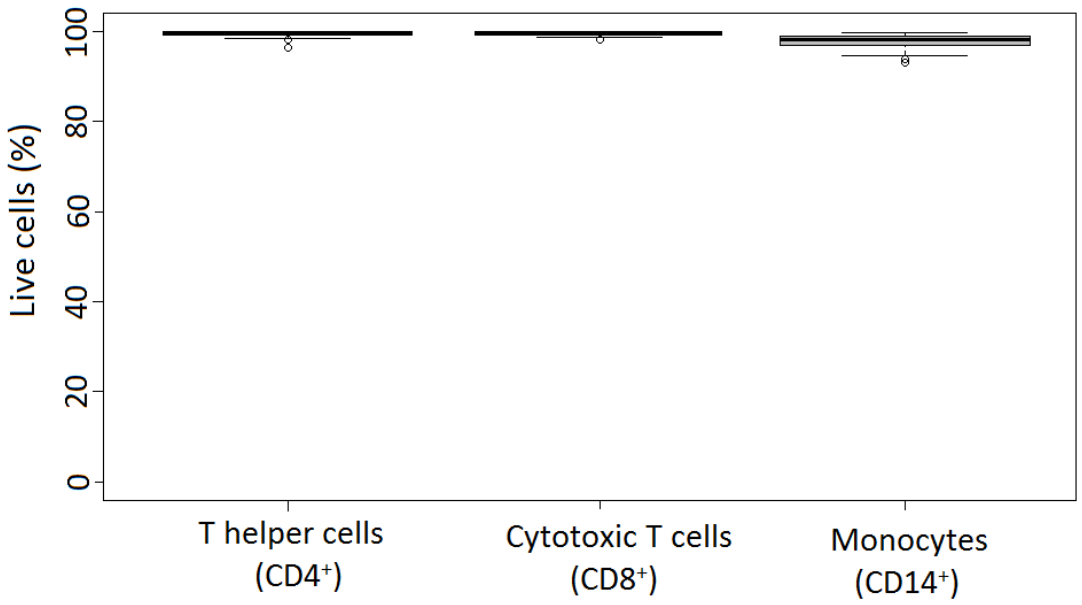


Figure 8.2 | Cell purities achieved post cell sorting
Box and whiskers plot visualising collective flow cytometry data plotted as percentage (%) live cells of $CD4^{pos}$ T helper cells, $CD8^{pos}$ cytotoxic T cells and $CD14^{pos}$ monocytes cell population post cell sorting, indicating cell purity. The box representing the 25^{th} and 75^{th} percentile from the median and the whiskers repressing the lowest and highest value. n=55.

Post cell sorting, between $4.7x10^5$ and $3.3x10^6$ purified cells were obtained with an average of $8.2x10^5$ purified cells per cell type. One patient sample was lost during cell sorting and thus was excluded from further analysis.

8.3 Immune cell stimulation and RNA quality control

Previous studies suggest that regulatory changes in gene expression in subsets of immune cells, and activation of immune-enhancers are more apparent upon immune stimulation, it was therefore decided to stimulate a portion of the

separated immune cell subsets. All separated immune cell subsets for which more than $2x10^5$ cells were obtained, were divided in two equal parts prior to a 4-hour incubation at 37°C and 5% CO_2. Half of the cells were left unstimulated and the other half were incubated with stimulatory agent. $CD4^{pos}$ and $CD8^{pos}$ cells were activated using $CD3^{pos}/CD28^{pos}$ T-activator beads and $CD14^{pos}$ cells were activated using LPS (lipopolysaccharides). Stimulation was assessed by relative quantification (RQ) of the TNFα gene by qPCR (see Materials and Methods 2.3.7). A 10-15-fold increase was observed in the $CD8^{pos}$ and $CD14^{pos}$ cells, where $CD4^{pos}$ cells showed a 98-fold increase of TNFα expression within the activated cells versus unstimulated cells (**Figure 8. 3**).

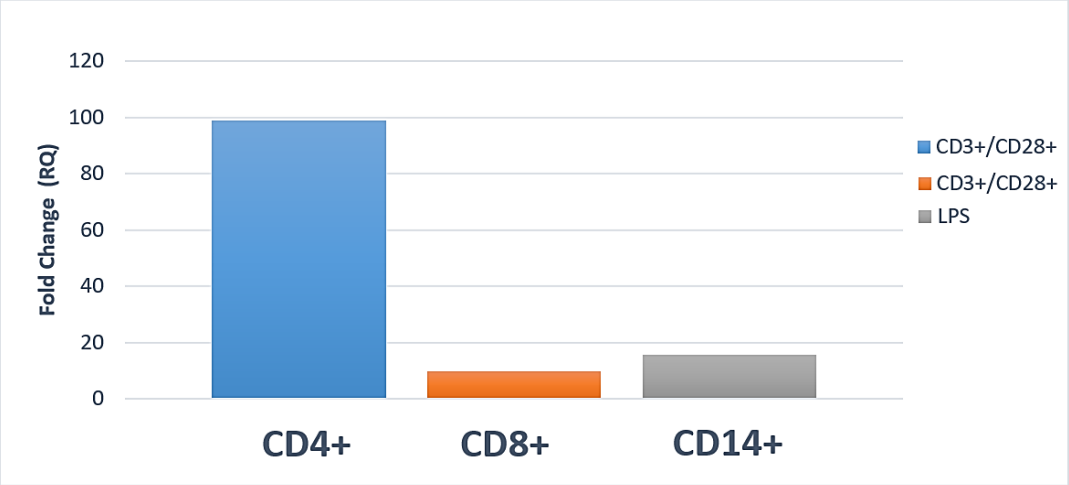


Figure 8. 3 | TNFα expression post stimulation
Fold-change/increase in expression of TNFα following stimulation; $CD3^+/CD28^+$ (blue and red) or LPS (green) of the isolated $CD4^{pos}$ T helper cells, $CD8^{pos}$ cytotoxic T cells, and $CD14^{pos}$ monocytes. (n=1)

Between 13 ng and 14.3 μg, with an average of 159 ng, of total RNA was extracted from separated unstimulated and activated $CD4^{pos}$, $CD8^{pos}$ and $CD14^{pos}$ cells, with RNA integrity and quality observed to vary but overall to be adequate to high (**Figure 8.5**).

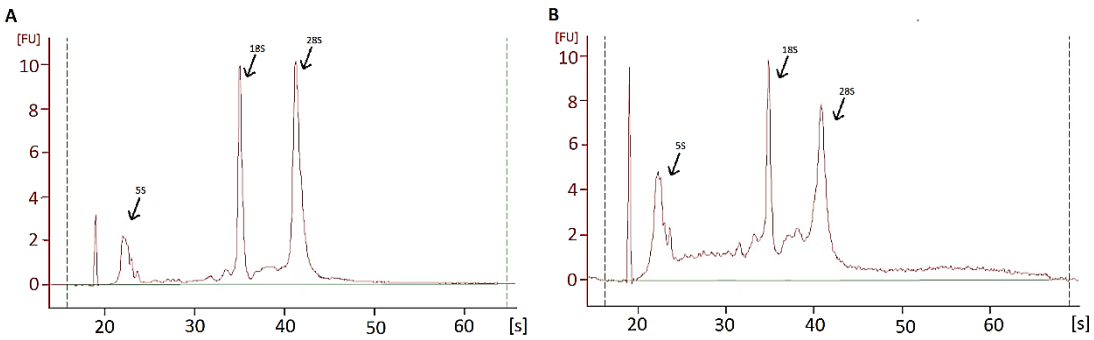


Figure 8.4 | Quality control RNA
Electropherograms displaying fluorescence units (FU) on y axis versus time (S), on the x axis. Showing marker (28sec), small RNAs <200nt (23sec), 18s subunit (~35sec) and 28s subunit (~42sec). (**A**) High quality sample with no degradation, (**B**) RNA showing an minor level of degradation. (Plots generated through the bioanalyzer (Agilent Technologies)).

No RIN scores were generated but based on electropherograms it was observed that the majority (82%) showed no or minor degradation, indicated by the absence of fluorescence between the small RNA and 18s peak at 24-30 sec (**Figure 8.5**). Approximately, 13% of RNA samples showed moderate degradation and 4% showed major degradation. Two samples were observed to be fully degraded and thus were omitted for the study. For this study 162 unstimulated - 54 samples x 3 immune cell types – and 109 stimulated RNA samples were taken forward for amplification and labelling for microarray.

8.4 Amplification and labelling

Amplification and labelling of the RNA samples was performed by Dr David Chambers, a lecturer in function genomics the CARD (Centre for Age Related Diseases). Amplification was performed using 5 ng of total RNA for each of the 271 RNA samples, producing 2-4 μg of cDNA per sample. Following successful amplification, the samples were biotinylated producing single stranded labelled cDNA ready for hybridisation to the Illumina HT-12 expression bead array. Quality of quantification and labelling was confirmed using the Agilent bioanalyzer (**Figure 8.6**).

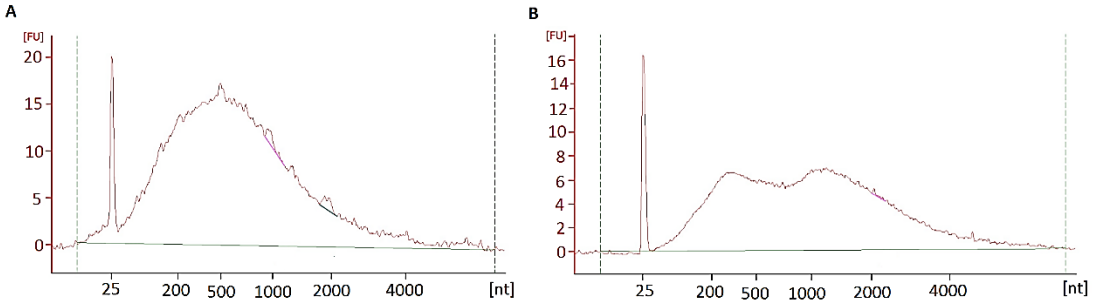


Figure 8.5 | Quality control
Electropherograms showing fluorescence units (FU) on y axis versus size, nucleotides (nt), on the x axis **(A)** Single stranded cDNA (100-2000 nt) and markers following labelling using Encore BiotinIL kit **(B)** Single stranded cDNA (100-2000 nt) and markers following labelling using Encore BiotinIL kit. (Plots generated through the bioanalyzer (Agilent Technologies)).

A symmetric peak sized between 100 and 2000 nucleotides was observed for 2 out of 58 the samples (**Figure 8.6A**), whereas the remainder of the samples showed a more irregular shapes as indicated by **Figure 8.6B**. Levels of fluorescence also varied widely from 2 FU up to 15 FU at the highest point of the peak.

8.5 Microarray results

The samples were normalised to 150 ng/μl of labelled single stranded cDNA and provided to the BRC Genomics Unit for processing using the Illumina HT-12 expression bead chip. The first 96 samples were processed, and the results assessed using Genomestudio. Extremely low background signal levels were observed within 96 microarray samples. The background signal is generated by a subset of probes of random sequence without any target in the genome. The expected values for background signal are around 100 to 150 strength of signal, however a much reduced background signal range of 0 – 50 was observed within the 96 microarray samples. For probes corresponding to genes, a signal range between 40-160 was detected; also lower than the average expected value of approximately 200. A clear batch affect was present with the first 48 samples, demonstrating detectable expression levels at $p = 0.05$ for an average 5,500 genes, compared to the average of detected genes at $p = 0.05$ for the second 48 samples, which was 2,000 (**Figure 8.7**). Due to the low background

signal observed, the low number of genes detected and the apparent batch effect, it was advised to discontinue the microarray experiments at this stage, until a potential cause for this poor quality data could be identified.

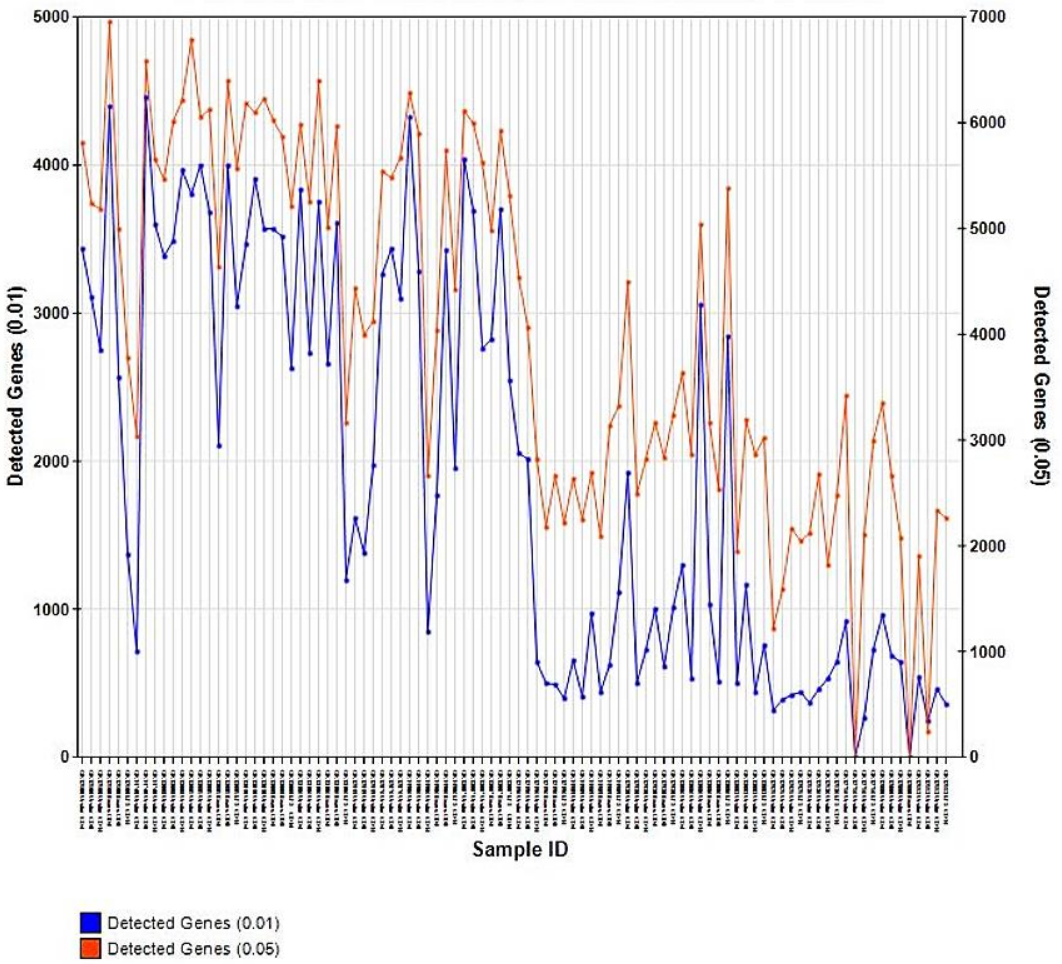


Figure 8.6 | Detected signal above background 96 samples
Number of genes for which expression was detected above background p=0.05 (orange line) and p=0.01 (blue line) per sample. (Plot generated by GenomeStudio, Illumina).

8.5.1 Troubleshooting

The unexpectedly low background and gene expression signals recorded in the microarray experiment could have potentially been a result of inaccurate starting quantification of samples, or poor quality input RNA. Re-quantification of the samples using Qubit fluorometric quantitation technology confirmed

original values to be correct, suggesting this was not the cause of the low expression signal observed. Furthermore, the quality of the RNA samples was reviewed by bioanalyzer traces generated using the labelled single stranded cDNA (sscDNA) (**Figure 8.6**). Although peaks were observed for all samples and suggested that labelled product of the appropriate size was achieved, the majority of sample traces did not conform to the recommended symmetrical size distribution required for high quality expression data (see Figure 8.6A for example). Only 2 samples out of 96 tested were consistent with the recommended high quality bioanalyzer profile, whereas the remaining sample profiles were highly variable as shown by Figure 8.6B. It was therefore suspected that the amplification and/or labelling of the RNA samples had been sub-optimal. The most likely cause is partial degradation of the RNA samples. Considering that the amplification kit relies on 3' poly-A tail binding, any level of degradation within the RNA will affect the amplification process. Having said this, RNA quality was checked using the bioanalyzer (**Figure 8.5**) and the majority of RNA samples showed adequate to high quality RNA, with 82% showing no to minor degradation. Another potential cause could be errors with the labelling kit or procedure.

In order to utilise these samples and generate data from them, a pilot set of 8 samples including $CD14^{pos}$ monocytes stimulated and non-stimulated, from two patients maintaining remission and two exhibiting relapse, were submitted of for mRNA sequencing.

8.6 Discussion

The aim of this study was to identify a panel of new and effective biomarkers for the prediction of likely relapse in patients suffering from CD. Relapse is often subclinical initially, but even at such a stage can be associated with irreversible bowel damage. The ability to predict which patients are more likely to relapse will allow the targeting of expensive drugs, restricted in their availability, to the appropriate patients as well as avoiding exposing patients to unnecessary medical therapy with potentially serious side effects. An important

factor to consider when investigating transcriptional biomarkers for clinical use is the selection of appropriate cells or tissue. Although, PBMCs are an easy tissue to obtain they are a heterogeneous cell population which has been shown to reduce the ability to pick up transcriptional signals [264]. Within IBD research, Lee et all has shown that the separation into purified cell subtypes can strengthen a transcriptional signal. They identified a transcriptional signature associated with a more severe disease progression using isolated $CD8^{pos}$ cells, but failed to observe the same signal in whole PBMCs [144]. It was therefore decided to look in purified subpopulations of peripheral blood, $CD4^{pos}$ T helper cells, $CD8^{pos}$ cytotoxic T cells and $CD14^{pos}$ monocytes, to investigate biomarkers for relapse in CD. These specific immune cell subtypes were chosen as they play important roles in cell-mediated immunity and have been shown to have a key role in the inflammatory response in IBD; $CD4^{pos}$ T cells are crucial in the pathogenesis of CD as they represent the majority of the activated mononuclear cells that infiltrate the intestinal wall [265]. $CD8^{pos}$ cells can be found in the mucosa in mouse models of IBD and several studies have suggested that autoreactive $CD8^{pos}$ T cells may be involved in the initiation of the inflammatory response in IBD [266]. $CD14^{pos}$ monocytes respond to microbial products such as LPS and Th1-derived IFN-gamma and are important in mucosal immunity [267].

$CD4^{pos}$ T helper cells, $CD8^{pos}$ cytotoxic T cells and $CD14^{pos}$ monocytes, from 49 patients in remission at time of entry into the study, were successfully purified, stimulated and RNA was extracted that was of relatively low yield. The amplification, labelling and hybridisation of the samples to the gene expression microarrays (HumanHT-12, Illumina) was outsourced to a local genomics facility, however, they were unable to generate data of sufficient quality from the first 96 samples. Troubleshooting indicated that the most likely reason for this was that the quality of the extracted RNA was more compromised than realised or an error was made in the amplification and labelling process of the samples. The employed amplification and labelling chemistries are highly sensitive to degradation of the RNA poly-A tail. Considering the limited amount

of remaining RNA, it was decided to not repeat the microarray analysis but to investigate an alternative method for gene expression profiling using low-input RNA sequencing. The initial quality control (QC) of the eight samples submitted for RNA sequencing indicated high quality data, with data analysis currently being performed. The hope is that this project, ones completed, would addresses an important unmet clinical need.

9. Conclusions and Future directions

9.1 Conclusions

This study aimed to increase the knowledge of the pathogenesis of IBD by characterising the entire colonic transcriptome, using whole RNA sequencing, and investigate gene expression at IBD susceptibility loci in biological relevant intestinal tissue from affected patients and controls. Following successful RNA sequencing of large intestinal tissue, differential expression between cases and controls was assessed, affected canonical pathways were identified and correlations between changes in gene expression and GWAS index SNPs in IBD were established. Heterogeneity of the intestinal biopsies used to generate the RNA sequencing data was assessed through cellular phenotyping. Finally, a pilot study was performed to investigate the transcriptional signature within multiple peripheral immune cells to predict relapse in CD patient.

Our study was the first to quantify the whole human transcriptome within uninflamed large intestinal tissue in IBD patient and controls, enabling hypothesis free quantification of coding and non-coding transcripts. It was shown that only 32% of the transcriptome exhibited expression above background within uninflamed IBD disease relevant tissue. Furthermore, it was established that 2,971 transcripts mapped to known IBD susceptibility loci, allowing us to prioritise genes potentially involved in the pathogenesis of CD, IBD or UC.

Differential expression analysis on the various IBD sub-phenotypes overall identified 1,637 genes exhibiting significant differences in expression within large intestinal tissue. Of these, 284 genes were prioritised to potentially be involved in the pathogenesis of IBD, UC or CD based on their genomic location within a known IBD susceptibility loci. The gene with the highest significance, showing reduced expression within CD and IBD cases *versus* controls, was *GLS* (Glutaminase). Glutaminase is involved in the breakdown of glutamine into glutamate and ammonia. Glutamine provides and important energy source for cells and has been shown to be essential to immune function

in the gut. Furthermore, reduced glutamine can lead to reduced gut mucosal integrity and increased gut permeability to allergens and pathogens. The role of glutaminase in the gut is less defined, although it has previously been observed that both intestinal glutamine levels and glutaminase activity are reduced in CD patients. The identified significant reduction in *GLS* expression in CD and IBD patients, confirmed previously reported results and suggests that both glutamine levels as well as breakdown of glutamine by glutaminase might contribute to IBD pathogenesis. Another gene highlighted as a strong candidate gene to be involved in IBD pathogenesis was *GAL3ST2*. *GAL3ST2* (Galactose-3-O-Sulfotransferase 2), exhibited a lower expression in UC vs CD patients. GAL3ST2 is known to be present in intestinal mucosa where it is involved in the synthesis of sulfomucins; sulfomucins have been implicated in the protection of the intestinal mucosa through increased mucus viscosity. Previous studies have shown that a significant loss of sulfomucins can be observed in the mucosal lining of UC patients. It is possible this loss of sulfomucins could be in some part due to the observed reduced colonic expression of *GAL3ST2*. Further investigation into this pathway is warranted in our cohort.

In addition to investigating individual genes exhibiting differential expression, we utilised pathway analysis tools to investigate underlying biological pathway affected by the differential expressed genes and drive hypothesis about biological pathways underlying disease pathology or etiology. Overall, 49 biological pathways were implicated in CD, IBD or UC. The majority of the identified pathways were involved in processes known to play an important role in IBD: immunoregulatory, autophagy and transmembrane signalling. The most significant perturbed pathways fell within three major groups: Gas and G-protein signalling pathways, the Notch signalling pathway and drug metabolism. One novel finding was the perturbation of Nicotine degradation pathway II and III within CD patients *versus* controls and UC patients. Smoking has been shown to have strongly opposing effects on the clinical course of IBD subtypes; smoking has proven to be beneficial to clinical remission in UC patients whereas, CD patients report detrimental effects. Considering nicotine

is an important component in smoking, our finding of two nicotine degradation pathways being perturbed solely within CD patients is interesting. Although, a multitude of research studies have been performed, an explanation for the opposing effects of smoking on UC and CD has not yet been found. The perturbation of nicotine degradation pathways II and III in CD patients and not UC patients, might be the first indication into the underlying mechanism of these differences in clinical outcomes in response to smoking.

Expression quantitative trait (eQTL) analysis is a powerful tool able to generate insights into associations between SNPs and changes in gene expression. Within complex diseases, eQTL analysis has proven highly valuable considering it can be employed to combine GWAS disease specific associations, often located in non-coding regions, with functional knowledge. This way eQTL results can generate valuable insights into disease mechanism and pathogenesis. Our eQTL analysis identified 126 *cis*-eQTLs located with known IBD loci, of these 23 were previously reported within colonic tissue. Furthermore, expression of 9 genes located within an IBD loci showed association with an IBD risk SNPs, making them strong candidate genes in IBD pathogenesis and further investigation should be performed. One of these 9 was *FAM49B* (Family With Sequence Similarity 49, member B), a gene which exhibited a 2-fold increased expression in the presence of the minor allele of IBD risk SNP rs13340584. In addition, *FAM49B* was previously prioritised to be involved in IBD pathogenesis and was observed to exhibit significant differences in expression between IBD cases and controls. Limited functional knowledge is available for FAM49B, although one study suggested a potential role in antigenic peptide presentation on a subset of T cells in the absence of ERAAP, with ERAAP1 and ERAAP2 also having been identified as *cis*-eQTLs with IBD risk SNPs. Overall, the eQTL analysis has contributed to our knowledge of IBD pathogenesis and the role of aberrant gene regulation within IBD, although additional investigation will be required.

Furthermore, our research attempted to address an important issue in RNA sequencing: tissue heterogeneity. A deconvolution model was build based on gene expression and cellular phenotypes of the intestinal biopsies. The developed deconvolution model showed for the first time the ability to predict cell fractions with a 100% accuracy (in the training set) without known 'marker genes' or epigenetic markers of specific cell types in the heterogeneous tissue. Unfortunately, when applied to a greater subset of biopsies the method proved inaccurate in its predictions. Contributing factors to this include the small sample size of the training set, cell viability, potential RNAseq batch effects or variation in the sites from where the colonic biopsies were taken e.g. transverse or descending colon. Although further optimisation is needed, progress towards the deconvolution of heterogeneous tissues was made.

Finally, as a separate project, we aimed to investigate biomarkers predictive of relapse in CD. Transcriptional profiles were attempted to be generated from immune cell subtypes within peripheral blood; $CD4^{pos}$ T helper cells, $CD8^{pos}$ cytotoxic T cells and $CD14^{pos}$ monocytes, within patients at time of remission and clinical relapse. These specific immune cell subtypes were chosen as they play important roles in cell-mediated immunity and have been shown to have a key role in the inflammatory response in IBD. Unfortunately, quantification of expression through microarray analysis was unsuccessful and no downstream analysis has yet been performed. Preliminary data from RNA sequencing looked promising and results will be expected in the future. This project, although not yet complete, aims to address an important unmet clinical need.

9.2 Future directions

When investigating genes exhibiting differences in colonic gene expression between IBD sub-phenotypes and controls, the UC vs control analysis generated p-values which plateaued out, most likely due to limited power with 24 UC patients and 28 controls. Although, it was possible to identify various genes suggested to affect UC pathogenesis using the larger CD group as a comparator through the UC vs CD analysis (24 vs 76), the number of UC patient samples will need to be increased to enable a more powerful UC vs control differential expression analysis. The identification of genes differentially expressed between UC vs control intestinal tissue, combined with the already performed UC vs CD analysis could provide valuable insights into genes contributing specifically to UC pathogenesis. Although, there is a clear overlap in disease features and biological pathways underlying UC and CD, investigation into their sub-phenotype specific etiology should also be performed.

The expression quantitative trait (eQTL) analysis performed in our study has provided additional insight into associations between SNPs covering 118 known IBD susceptibility loci and changes in gene expression of nearby genes. Unfortunately, 106 out of 224 IBD susceptibility loci were not covered by our genotype data. Through imputation of SNP alleles at these un-genotyped IBD risk loci using 1000 genomes data, further insights into genes affected within these 106 IBD susceptibility loci could be generated. Through the investigation of GWAS association signals with functional studies, such as eQTL studies, insights into the mechanistic etiology of complex diseases can potentially be generated.

Although good progress has been made in identifying the effect of IBD susceptibility SNPs on proximal or nearby genes, it is hypothesised that a subset of the IBD causal SNPs will result in changes in expression of more distal genes, outside the 1Mb that was tested in this study, or even on different chromosomes. *Trans*-eQTL studies are limited in their power due to the numbed of association tests required. To perform a well powered *trans*-eQTL

analysis, a samples size of > 600 samples would ideally be achieved. With RNA sequencing datasets more frequently being made publically available, a potential future meta-analysis might provide the power needed for a *trans*-eQTL analysis in uninflamed intestinal tissue of IBD patients.

The identification of cell type specific expression signals within heterogeneous tissues will need to be addressed in future research. This can either be done through optimisation of our deconvolution method or by purifying cell populations using flow cytometry. Purifying cell populations using flow cytometry has become more appealing with low input or single cell RNA sequencing technologies now being accessible, although it remains to be a labour intensive and expensive method which might affect the transcriptome. Even though, the by us developed method for deconvolution of the intestinal biopsy samples was not accurate in predicting the cell type fractions within our intestinal biopsies, it showed major promise by achieving a 100% accuracy in the biopsy data set used to build the model. Optimisation of the deconvolution model should include expansion of biopsy cellular phenotype data (well matched to the colonic site of where they were taken) and incorporation of RNAseq batch effects into the model. If the deconvolution model can be optimised, the generated data for all 127 patients and controls can be further utilised to investigate if observed differences in expression are caused by phenotype or variation in biopsy composition. Furthermore, accurate biopsy composition predictions could potentially enable identification of tissue specific eQTLs.

Although, whole RNA sequencing was performed on the intestinal biopsies this did not include micro RNAs (miRNAs). miRNAs are known to be of great importance in transcriptional regulation through RNA silencing and post-transcriptional regulation of gene expression. Furthermore, miRNAs have been reported to play a key role in the regulation of immune development cells and aberrant expression of certain miRNAs could contribute to autoimmunity [268]. Considering IBD susceptibility SNPs are known to be located in between genes

and hypothesised to affect genes through an indirect fashion, miRNAs and their role in IBD will need to be investigated. Our aim is to sequence all colonic miRNAs within the 127 intestinal biopsies collected for the whole RNA sequencing study and perform a similar analysis package i.e. investigate differential expression between sub phenotypes, identify underlying biological pathways and incorporate the miRNAs into the eQTL analysis.

Top hits from both the miRNA and whole RNA sequencing result should be validated using real-time qPCR.

Finally, preliminary quality control (QC) has shown that the eight samples submitted for RNA sequencing, in the biomarkers for relapse study, generated high quality data. This confirms that the RNA quality of the collected samples can be sufficiently high to generate good quality transcriptional profiles. The remaining 263 unstimulated and stimulated $CD4^{pos}$ and $CD8^{pos}$ T cells and $CD14^{pos}$ monocytes should therefore also be submitted for RNA sequencing. Comparison of the transcriptional profiles from patients identified as relapsed to non-relapsed patients as well as stimulated vs unstimulated samples will aim to identify biomarkers predictive of likely relapse in CD patients.

References

1. Crohn BB, Ginzburg L, Oppenheimer GD. Regional ileitis. A pathological and clinical entity. *JAMA*. 1932;99(16):1323-1329.

2. Lichtenstein GR, Hanauer SB, Sandborn WJ, Practice Parameters Committee of American College of Gastroenterology. Management of crohn's disease in adults. *Am J Gastroenterol*. 2009;104(2):465-83; quiz 464, 484.

3. Baumgart DC, Sandborn WJ. Crohn's disease. *Lancet*. 2012;380(9853):1590-1605.

4. Jung C, Hugot JP, Barreau F. Peyer's patches: The immune sensors of the intestine. *Int J Inflam*. 2010;2010:823710.

5. Carbonnel F, Macaigne G, Beaugerie L, Gendre JP, Cosnes J. Crohn's disease severity in familial and sporadic cases. *Gut*. 1999;44(1):91-95.

6. Scherr R, Essers J, Hakonarson H, Kugathasan S. Genetic determinants of pediatric inflammatory bowel disease: Is age of onset genetically determined? *Dig Dis*. 2009;27(3):236-239.

7. Gardiner KR, Dasari BV. Operative management of small bowel crohn's disease. *Surg Clin North Am*. 2007;87(3):587-610.

8. Wilkens S. Morbid appearances in the intestine of Miss Bankes. London Medical Times & Gezette. 1985;2;264.

9. Ordas I, Eckmann L, Talamini M, Baumgart DC, Sandborn WJ. Ulcerative colitis. *Lancet*. 2012;380(9853):1606-1619.

10. Carter MJ, Lobo AJ, Travis SP, IBD Section, British Society of Gastroenterology. Guidelines for the management of inflammatory bowel disease in adults. *Gut*. 2004;53 Suppl 5:V1-16.

11. Dendrinos K, Cerda S, Farraye FA. The "cecal patch" in patients with ulcerative colitis. *Gastrointest Endosc*. 2008;68(5):1006-7; discussion 1007.

12. Loftus EV,Jr, Sandborn WJ. Epidemiology of inflammatory bowel disease. *Gastroenterol Clin North Am*. 2002;31(1):1-20.

13. Danese S, Vuitton L, Peyrin-Biroulet L. Biologic agents for IBD: Practical insights. *Nat Rev Gastroenterol Hepatol*. 2015;12(9):537-545.

14. Chaparro M, Panes J, Garcia V, et al. Long-term durability of infliximab treatment in crohn's disease and efficacy of dose "escalation" in patients losing response. *J Clin Gastroenterol*. 2011;45(2):113-118.

15. Eshuis EJ, Peters CP, van Bodegraven AA, et al. Ten years of infliximab for crohn's disease: Outcome in 469 patients from 2 tertiary referral centers. *Inflamm Bowel Dis*. 2013;19(8):1622-1630.

16. Reinisch W, Sandborn WJ, Rutgeerts P, et al. Long-term infliximab maintenance therapy for ulcerative colitis: The ACT-1 and -2 extension studies. *Inflamm Bowel Dis*. 2012;18(2):201-211.

17. Sandborn WJ, Feagan BG, Rutgeerts P, et al. Vedolizumab as induction and maintenance therapy for crohn's disease. *N Engl J Med*. 2013;369(8):711-721.

18. Feagan BG, Rutgeerts P, Sands BE, et al. Vedolizumab as induction and maintenance therapy for ulcerative colitis. *N Engl J Med*. 2013;369(8):699-710.

19. Sandborn WJ, Feagan BG, Rutgeerts P, et al. Vedolizumab as induction and maintenance therapy for crohn's disease. *N Engl J Med*. 2013;369(8):711-721.

20. Feagan BG, Sandborn WJ, Gasink C, et al. Ustekinumab as induction and maintenance therapy for crohn's disease. *N Engl J Med*. 2016;375(20):1946-1960.

21. Singh JA, Wells GA, Christensen R, et al. Adverse effects of biologics: A network meta-analysis and cochrane overview. *Cochrane Database Syst Rev*. 2011;(2):CD008794. doi(2):CD008794.

22. Bodger K, Kikuchi T, Hughes D. Cost-effectiveness of biological therapy for crohn's disease: Markov cohort analyses incorporating united kingdom patient-level cost data. *Aliment Pharmacol Ther*. 2009;30(3):265-274.

23. van der Valk ME, Mangen MJ, Leenders M, et al. Healthcare costs of inflammatory bowel disease have shifted from hospitalisation and surgery towards anti-TNFalpha therapy: Results from the COIN study. *Gut*. 2014;63(1):72-79.

24. Bernstein CN, Ng SC, Lakatos PL, Moum B, Loftus EV,Jr, Epidemiology and Natural History Task Force of the International Organization of the Study of Inflammatory Bowel Disease. A review of mortality and surgery in ulcerative colitis: Milestones of the seriousness of the disease. *Inflamm Bowel Dis*. 2013;19(9):2001-2010.

25. Molodecky NA, Soon IS, Rabi DM, et al. Increasing incidence and prevalence of the inflammatory bowel diseases with time, based on systematic review. *Gastroenterology*. 2012;142(1):46-54.e42; quiz e30.

26. Thompson NP, Fleming DM, Charlton J, Pounder RE, Wakefield AJ. Patients consulting with crohn's disease in primary care in england and wales. *Eur J Gastroenterol Hepatol*. 1998;10(12):1007-1012.

27. Rubin GP, Hungin AP, Kelly PJ, Ling J. Inflammatory bowel disease: Epidemiology and management in an english general practice population. *Aliment Pharmacol Ther*. 2000;14(12):1553-1559.

28. Vind I, Riis L, Jess T, et al. Increasing incidences of inflammatory bowel disease and decreasing surgery rates in copenhagen city and county, 2003-2005: A population-based study from the danish crohn colitis database. *Am J Gastroenterol*. 2006;101(6):1274-1282.

29. Lapidus A. Crohn's disease in stockholm county during 1990-2001: An epidemiological update. *World J Gastroenterol*. 2006;12(1):75-81.

30. Busch K, Ludvigsson JF, Ekstrom-Smedby K, Ekbom A, Askling J, Neovius M. Nationwide prevalence of inflammatory bowel disease in sweden: A population-based register study. *Aliment Pharmacol Ther*. 2014;39(1):57-68.

31. Burisch J, Jess T, Martinato M, Lakatos PL, ECCO -EpiCom. The burden of inflammatory bowel disease in europe. *J Crohns Colitis*. 2013;7(4):322-337.

32. Shivananda S, Lennard-Jones J, Logan R, et al. Incidence of inflammatory bowel disease across europe: Is there a difference between north and south? results of the european collaborative study on inflammatory bowel disease (EC-IBD). *Gut*. 1996;39(5):690-697.

33. Betteridge JD, Armbruster SP, Maydonovitch C, Veerappan GR. Inflammatory bowel disease prevalence by age, gender, race, and geographic location in the U.S. military health care population. *Inflamm Bowel Dis*. 2013;19(7):1421-1427.

34. Tysk C, Lindberg E, Jarnerot G, Floderus-Myrhed B. Ulcerative colitis and crohn's disease in an unselected population of monozygotic and dizygotic twins. A study of heritability and the influence of smoking. *Gut*. 1988;29(7):990-996.

35. Spehlmann ME, Begun AZ, Burghardt J, Lepage P, Raedler A, Schreiber S. Epidemiology of inflammatory bowel disease in a german twin cohort: Results of a nationwide study. *Inflamm Bowel Dis*. 2008;14(7):968-976.

36. Probert CS, Jayanthi V, Hughes AO, Thompson JR, Wicks AC, Mayberry JF. Prevalence and family risk of ulcerative colitis and crohn's disease: An epidemiological study among europeans and south asians in leicestershire. *Gut*. 1993;34(11):1547-1551.

37. Peeters M, Nevens H, Baert F, et al. Familial aggregation in crohn's disease: Increased age-adjusted risk and concordance in clinical characteristics. *Gastroenterology*. 1996;111(3):597-603.

38. Halme L, Paavola-Sakki P, Turunen U, Lappalainen M, Farkkila M, Kontula K. Family and twin studies in inflammatory bowel disease. *World J Gastroenterol*. 2006;12(23):3668-3672.

39. Lewis CM, Knight J. Introduction to genetic association studies. *Cold Spring Harb Protoc*. 2012;2012(3):297-306.

40. Chen GB, Lee SH, Brion MJ, et al. Estimation and partitioning of (co)heritability of inflammatory bowel disease from GWAS and immunochip data. *Hum Mol Genet*. 2014;23(17):4710-4720.

41. Bridger S, Lee JC, Bjarnason I, Jones JE, Macpherson AJ. In siblings with similar genetic susceptibility for inflammatory bowel disease, smokers tend to develop crohn's disease and non-smokers develop ulcerative colitis. *Gut*. 2002;51(1):21-25.

42. Mahid SS, Minor KS, Soto RE, Hornung CA, Galandiuk S. Smoking and inflammatory bowel disease: A meta-analysis. *Mayo Clin Proc*. 2006;81(11):1462-1471.

43. Birrenbach T, Bocker U. Inflammatory bowel disease and smoking: A review of epidemiology, pathophysiology, and therapeutic implications. *Inflamm Bowel Dis*. 2004;10(6):848-859.

44. Persson PG, Hellers G, Ahlbom A. Use of oral moist snuff and inflammatory bowel disease. *Int J Epidemiol*. 1993;22(6):1101-1103.

45. Barton JR, Riad MA, Gaze MN, Maran AG, Ferguson A. Mucosal immunodeficiency in smokers, and in patients with epithelial head and neck tumours. *Gut*. 1990;31(4):378-382.

46. Srivastava ED, Barton JR, O'Mahony S, et al. Smoking, humoral immunity, and ulcerative colitis. *Gut*. 1991;32(9):1016-1019.

47. Sher ME, Bank S, Greenberg R, et al. The influence of cigarette smoking on cytokine levels in patients with inflammatory bowel disease. *Inflamm Bowel Dis*. 1999;5(2):73-78.

48. Bergeron V, Grondin V, Rajca S, et al. Current smoking differentially affects blood mononuclear cells from patients with crohn's disease and ulcerative colitis: Relevance to its adverse role in the disease. *Inflamm Bowel Dis*. 2012;18(6):1101-1111.

49. Kalra J, Chaudhary AK, Prasad K. Increased production of oxygen free radicals in cigarette smokers. *Int J Exp Pathol*. 1991;72(1):1-7.

50. Scott AM, Kellow JE, Eckersley GM, Nolan JM, Jones MP. Cigarette smoking and nicotine delay postprandial mouth-cecum transit time. *Dig Dis Sci*. 1992;37(10):1544-1547.

51. McGilligan VE, Wallace JM, Heavey PM, Ridley DL, Rowland IR. Hypothesis about mechanisms through which nicotine might exert its effect on the interdependence of inflammation and gut barrier function in ulcerative colitis. *Inflamm Bowel Dis*. 2007;13(1):108-115.

52. Nielsen OH, Bjerrum JT, Csillag C, Nielsen FC, Olsen J. Influence of smoking on colonic gene expression profile in crohn's disease. *PLoS One*. 2009;4(7):e6210.

53. Andersson RE, Olaison G, Tysk C, Ekbom A. Appendectomy and protection against ulcerative colitis. *N Engl J Med*. 2001;344(11):808-814.

54. Andersson RE, Olaison G, Tysk C, Ekbom A. Appendectomy is followed by increased risk of crohn's disease. *Gastroenterology*. 2003;124(1):40-46.

55. Ananthakrishnan AN, Khalili H, Konijeti GG, et al. A prospective study of long-term intake of dietary fiber and risk of crohn's disease and ulcerative colitis. *Gastroenterology*. 2013;145(5):970-977.

56. Devkota S, Wang Y, Musch MW, et al. Dietary-fat-induced taurocholic acid promotes pathobiont expansion and colitis in Il10-/- mice. *Nature*. 2012;487(7405):104-108.

57. Ananthakrishnan AN, Khalili H, Konijeti GG, et al. Long-term intake of dietary fat and risk of ulcerative colitis and crohn's disease. *Gut*. 2014;63(5):776-784.

58. Ananthakrishnan AN, Khalili H, Higuchi LM, et al. Higher predicted vitamin D status is associated with reduced risk of crohn's disease. *Gastroenterology*. 2012;142(3):482-489.

59. Ananthakrishnan AN, Cagan A, Gainer VS, et al. Normalization of plasma 25-hydroxy vitamin D is associated with reduced risk of surgery in crohn's disease. *Inflamm Bowel Dis*. 2013;19(9):1921-1927.

60. Lerebours E, Gower-Rousseau C, Merle V, et al. Stressful life events as a risk factor for inflammatory bowel disease onset: A population-based case-control study. *Am J Gastroenterol*. 2007;102(1):122-131.

61. Bernstein CN, Singh S, Graff LA, Walker JR, Miller N, Cheang M. A prospective population-based study of triggers of symptomatic flares in IBD. *Am J Gastroenterol*. 2010;105(9):1994-2002.

62. Bonaz BL, Bernstein CN. Brain-gut interactions in inflammatory bowel disease. *Gastroenterology*. 2013;144(1):36-49.

63. Sonnenberg A. Occupational distribution of inflammatory bowel disease among german employees. *Gut*. 1990;31(9):1037-1040.

64. Khalili H, Ananthakrishnan AN, Konijeti GG, et al. Physical activity and risk of inflammatory bowel disease: Prospective study from the nurses' health study cohorts. *BMJ*. 2013;347:f6633.

65. Ananthakrishnan AN, Long MD, Martin CF, Sandler RS, Kappelman MD. Sleep disturbance and risk of active disease in patients with crohn's disease and ulcerative colitis. *Clin Gastroenterol Hepatol*. 2013;11(8):965-971.

66. Weinstock JV, Elliott DE. Helminths and the IBD hygiene hypothesis. *Inflamm Bowel Dis*. 2009;15(1):128-133.

67. Rook GA, Adams V, Hunt J, Palmer R, Martinelli R, Brunet LR. Mycobacteria and other environmental organisms as immunomodulators for immunoregulatory disorders. *Springer Semin Immunopathol*. 2004;25(3-4):237-255.

68. Goodrich JK, Waters JL, Poole AC, et al. Human genetics shape the gut microbiome. *Cell*. 2014;159(4):789-799.

69. Brestoff JR, Artis D. Commensal bacteria at the interface of host metabolism and the immune system. *Nat Immunol*. 2013;14(7):676-684.

70. Eckburg PB, Bik EM, Bernstein CN, et al. Diversity of the human intestinal microbial flora. *Science*. 2005;308(5728):1635-1638.

71. Frank DN, St Amand AL, Feldman RA, Boedeker EC, Harpaz N, Pace NR. Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases. *Proc Natl Acad Sci U S A*. 2007;104(34):13780-13785.

72. Peterson DA, Frank DN, Pace NR, Gordon JI. Metagenomic approaches for defining the pathogenesis of inflammatory bowel diseases. *Cell Host Microbe*. 2008;3(6):417-427.

73. Martinez C, Antolin M, Santos J, et al. Unstable composition of the fecal microbiota in ulcerative colitis during clinical remission. *Am J Gastroenterol*. 2008;103(3):643-648.

74. Andoh A, Kuzuoka H, Tsujikawa T, et al. Multicenter analysis of fecal microbiota profiles in japanese patients with crohn's disease. *J Gastroenterol*. 2012;47(12):1298-1307.

75. Sokol H, Leducq V, Aschard H, et al. Fungal microbiota dysbiosis in IBD. *Gut*. 2017;66(6):1039-1048.

76. Norman JM, Handley SA, Baldridge MT, et al. Disease-specific alterations in the enteric virome in inflammatory bowel disease. *Cell*. 2015;160(3):447-460.

77. Xavier RJ, Podolsky DK. Unravelling the pathogenesis of inflammatory bowel disease. *Nature*. 2007;448(7152):427-434.

78. Heller F, Florian P, Bojarski C, et al. Interleukin-13 is the key effector Th2 cytokine in ulcerative colitis that affects epithelial tight junctions, apoptosis, and cell restitution. *Gastroenterology*. 2005;129(2):550-564.

79. Zeissig S, Burgel N, Gunzel D, et al. Changes in expression and distribution of claudin 2, 5 and 8 lead to discontinuous tight junctions and barrier dysfunction in active crohn's disease. *Gut*. 2007;56(1):61-72.

80. Wehkamp J, Harder J, Weichenthal M, et al. Inducible and constitutive beta-defensins are differentially expressed in crohn's disease and ulcerative colitis. *Inflamm Bowel Dis*. 2003;9(4):215-223.

81. Hart AL, Al-Hassi HO, Rigby RJ, et al. Characteristics of intestinal dendritic cells in inflammatory bowel diseases. *Gastroenterology*. 2005;129(1):50-65.

82. Cario E, Podolsky DK. Differential alteration in intestinal epithelial cell expression of toll-like receptor 3 (TLR3) and TLR4 in inflammatory bowel disease. *Infect Immun*. 2000;68(12):7010-7017.

83. Hugot JP, Chamaillard M, Zouali H, et al. Association of NOD2 leucine-rich repeat variants with susceptibility to crohn's disease. *Nature*. 2001;411(6837):599-603.

84. Baumgart DC, Metzke D, Guckelberger O, et al. Aberrant plasmacytoid dendritic cell distribution and function in patients with crohn's disease and ulcerative colitis. *Clin Exp Immunol*. 2011;166(1):46-54.

85. Fuss IJ, Neurath M, Boirivant M, et al. Disparate CD4+ lamina propria (LP) lymphokine secretion profiles in inflammatory bowel disease. crohn's disease LP cells manifest increased secretion of IFN-gamma, whereas ulcerative colitis LP cells manifest increased secretion of IL-5. *J Immunol*. 1996;157(3):1261-1270.

86. Gong Y, Lin Y, Zhao N, et al. The Th17/treg immune imbalance in ulcerative colitis disease in a chinese han population. *Mediators Inflamm*. 2016;2016:7089137.

87. Shale M, Schiering C, Powrie F. CD4(+) T-cell subsets in intestinal inflammation. *Immunol Rev*. 2013;252(1):164-182.

88. Cheroutre H. In IBD eight can come before four. *Gastroenterology*. 2006;131(2):667-670.

89. Mahida YR. The key role of macrophages in the immunopathogenesis of inflammatory bowel disease. *Inflamm Bowel Dis*. 2000;6(1):21-33.

90. Romagnani S. Lymphokine production by human T cells in disease states. *Annu Rev Immunol*. 1994;12:227-257.

91. Korn T, Bettelli E, Oukka M, Kuchroo VK. IL-17 and Th17 cells. *Annu Rev Immunol*. 2009;27:485-517.

92. Jostins L, Ripke S, Weersma RK, et al. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature*. 2012;491(7422):119-124.

93. Satsangi J, Parkes M, Louis E, et al. Two stage genome-wide search in inflammatory bowel disease provides evidence for susceptibility loci on chromosomes 3, 7 and 12. *Nat Genet*. 1996;14(2):199-202.

94. Hugot JP, Laurent-Puig P, Gower-Rousseau C, et al. Mapping of a susceptibility locus for crohn's disease on chromosome 16. *Nature*. 1996;379(6568):821-823.

95. Hampe J, Cuthbert A, Croucher PJ, et al. Association between insertion mutation in NOD2 gene and crohn's disease in german and british populations. *Lancet*. 2001;357(9272):1925-1928.

96. Ogura Y, Bonen DK, Inohara N, et al. A frameshift mutation in NOD2 associated with susceptibility to crohn's disease. *Nature*. 2001;411(6837):603-606.

97. Lesage S, Zouali H, Cezard JP, et al. CARD15/NOD2 mutational analysis and genotype-phenotype correlation in 612 patients with inflammatory bowel disease. *Am J Hum Genet*. 2002;70(4):845-857.

98. Rioux JD, Daly MJ, Silverberg MS, et al. Genetic variation in the 5q31 cytokine gene cluster confers susceptibility to crohn disease. *Nat Genet*. 2001;29(2):223-228.

99. Stoll M, Corneliussen B, Costello CM, et al. Genetic variation in DLG5 is associated with inflammatory bowel disease. *Nat Genet*. 2004;36(5):476-480.

100. Hampe J, Shaw SH, Saiz R, et al. Linkage of inflammatory bowel disease to human chromosome 6p. *Am J Hum Genet*. 1999;65(6):1647-1655.

101. Duerr RH, Taylor KD, Brant SR, et al. A genome-wide association study identifies IL23R as an inflammatory bowel disease gene. *Science*. 2006;314(5804):1461-1463.

102. Hampe J, Franke A, Rosenstiel P, et al. A genome-wide association scan of nonsynonymous SNPs identifies a susceptibility variant for crohn disease in ATG16L1. *Nat Genet*. 2007;39(2):207-211.

103. Libioulle C, Louis E, Hansoul S, et al. Novel crohn disease locus identified by genome-wide association maps to a gene desert on 5p13.1 and modulates expression of PTGER4. *PLoS Genet*. 2007;3(4):e58.

104. Rioux JD, Xavier RJ, Taylor KD, et al. Genome-wide association study identifies new susceptibility loci for crohn disease and implicates autophagy in disease pathogenesis. *Nat Genet*. 2007;39(5):596-604.

105. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*. 2007;447(7145):661-678.

106. Burdett T (EBI), Hall PN (NHGRI), Hastings E (EBI), Hindorff LA (NHGRI), Junkins HA (NHGRI), Klemm AK (NHGRI), MacArthur J (EBI), Manolio TA (NHGRI), Morales J (EBI), Parkinson H (EBI) and Welter D (EBI). The NHGRI-EBI Catalog of published genome-wide association studies. Available at: www.ebi.ac.uk/gwas. Accessed 03/05/2016, version 1.0.1.

107. Liu JZ, van Sommeren S, Huang H, et al. Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat Genet*. 2015;47(9):979-986.

108. de Lange KM, Moutsianas L, Lee JC, et al. Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. *Nat Genet*. 2017;49(2):256-261.

109. Raychaudhuri S, Plenge RM, Rossin EJ, et al. Identifying relationships among genomic disease regions: Predicting genes at pathogenic SNP associations and rare deletions. *PLoS Genet*. 2009;5(6):e1000534.

110. Rossin EJ, Lage K, Raychaudhuri S, et al. Proteins encoded in genomic regions associated with immune-mediated disease physically interact and suggest underlying biology. *PLoS Genet*. 2011;7(1):e1001273.

111. van de Bunt M, Cortes A, IGAS Consortium, Brown MA, Morris AP, McCarthy MI. Evaluating the performance of fine-mapping strategies at common variant GWAS loci. *PLoS Genet*. 2015;11(9):e1005535.

112. Huang H, Fang M, Jostins L, et al. Fine-mapping inflammatory bowel disease loci to single-variant resolution. *Nature*. 2017;547(7662):173-178.

113. Farh KK, Marson A, Zhu J, et al. Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature*. 2015;518(7539):337-343.

114. Prescott NJ, Dominy KM, Kubo M, et al. Independent and population-specific association of risk variants at the IRGM locus with crohn's disease. *Hum Mol Genet*. 2010;19(9):1828-1839.

115. Chauhan S, Mandell MA, Deretic V. Mechanism of action of the tuberculosis and crohn disease risk factor IRGM in autophagy. *Autophagy*. 2016;12(2):429-431.

116. Sivanesan D, Beauchamp C, Quinou C, et al. IL23R (interleukin 23 receptor) variants protective against inflammatory bowel diseases (IBD) display loss of function due to impaired protein stability and intracellular trafficking. *J Biol Chem*. 2016;291(16):8673-8685.

117. Ellinghaus D, Zhang H, Zeissig S, et al. Association between variants of PRDM1 and NDP52 and crohn's disease, based on exome sequencing and functional studies. *Gastroenterology*. 2013;145(2):339-347.

118. Granlund A, Flatberg A, Ostvik AE, et al. Whole genome gene expression meta-analysis of inflammatory bowel disease colon mucosa demonstrates lack of major differences between crohn's disease and ulcerative colitis. *PLoS One*. 2013;8(2):e56818.

119. Peloquin JM, Goel G, Kong L, et al. Characterization of candidate genes in inflammatory bowel disease-associated risk loci. *JCI Insight*. 2016;1(13):e87899.

120. Wu S, Zhang YG, Lu R, et al. Intestinal epithelial vitamin D receptor deletion leads to defective autophagy in colitis. *Gut*. 2015;64(7):1082-1094.

121. Kotka M, Lieden A, Pettersson S, Trinchieri V, Masci A, D'Amato M. Solute carriers (SLC) in inflammatory bowel disease: A potential target of probiotics? *J Clin Gastroenterol*. 2008;42 Suppl 3 Pt 1:S133-5.

122. Gaublomme JT, Yosef N, Lee Y, et al. Single-cell genomics unveils critical regulators of Th17 cell pathogenicity. *Cell*. 2015;163(6):1400-1412.

123. Stranger BE, Raj T. Genetics of human gene expression. *Curr Opin Genet Dev*. 2013;23(6):627-634.

124. Dixon AL, Liang L, Moffatt MF, et al. A genome-wide association study of global gene expression. *Nat Genet*. 2007;39(10):1202-1207.

125. Stranger BE, Nica AC, Forrest MS, et al. Population genomics of human gene expression. *Nat Genet*. 2007;39(10):1217-1224.

126. Montgomery SB, Sammeth M, Gutierrez-Arcelus M, et al. Transcriptome genetics using second generation sequencing in a caucasian population. *Nature*. 2010;464(7289):773-777.

127. http://www.gtexportal.org/home/tissueSummaryPage#sampleInfo. Accessed 04/05/2016, release V6.

128. GTEx Consortium. Human genomics. the genotype-tissue expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science*. 2015;348(6235):648-660.

129. Dimas AS, Deutsch S, Stranger BE, et al. Common regulatory variation impacts gene expression in a cell type-dependent manner. *Science*. 2009;325(5945):1246-1250.

130. Kabakchiev B, Silverberg MS. Expression quantitative trait loci analysis identifies associations between genotype and gene expression in human intestine. *Gastroenterology*. 2013;144(7):1488-96, 1496.e1-3.

131. Repnik K, Potocnik U. eQTL analysis links inflammatory bowel disease associated 1q21 locus to ECM1 gene. *J Appl Genet*. 2016.

132. Singh T, Levine AP, Smith PJ, Smith AM, Segal AW, Barrett JC. Characterization of expression quantitative trait loci in the human colon. *Inflamm Bowel Dis*. 2015;21(2):251-256.

133. Rahman A, Decourcey J, Larbi NB, Loughran ST, Walls D, Loscher CE. Syntaxin-4 is essential for IgE secretion by plasma cells. *Biochem Biophys Res Commun*. 2013;440(1):163-167.

134. Soubieres AA, Poullis A. Emerging role of novel biomarkers in the diagnosis of inflammatory bowel disease. *World J Gastrointest Pharmacol Ther*. 2016;7(1):41-50.

135. Darlington GJ, Wilson DR, Lachman LB. Monocyte-conditioned medium, interleukin-1, and tumour necrosis factor stimulate the acute phase response in human hepatoma cells in vitro. *J Cell Biol*. 1986;103(3):787-793.

136. Solem CA, Loftus EV,Jr, Tremaine WJ, Harmsen WS, Zinsmeister AR, Sandborn WJ. Correlation of C-reactive protein with clinical, endoscopic, histologic, and radiographic activity in inflammatory bowel disease. *Inflamm Bowel Dis*. 2005;11(8):707-712.

137. Joossens S, Reinisch W, Vermeire S, et al. The value of serologic markers in indeterminate colitis: A prospective follow-up study. *Gastroenterology*. 2002;122(5):1242-1247.

138. Reese GE, Constantinides VA, Simillis C, et al. Diagnostic precision of anti-saccharomyces cerevisiae antibodies and perinuclear antineutrophil cytoplasmic antibodies in inflammatory bowel disease. *Am J Gastroenterol*. 2006;101(10):2410-2422.

139. Poullis A, Foster R, Mendall MA, Fagerhol MK. Emerging role of calprotectin in gastroenterology. *J Gastroenterol Hepatol*. 2003;18(7):756-762.

140. Boussac M, Garin J. Calcium-dependent secretion in human neutrophils: A proteomic approach. *Electrophoresis*. 2000;21(3):665-672.

141. Wright EK, De Cruz P, Gearry R, Day AS, Kamm MA. Fecal biomarkers in the diagnosis and monitoring of crohn's disease. *Inflamm Bowel Dis*. 2014;20(9):1668-1677.

142. von Stein P, Lofberg R, Kuznetsov NV, et al. Multigene analysis can discriminate between ulcerative colitis, crohn's disease, and irritable bowel syndrome. *Gastroenterology*. 2008;134(7):1869-81; quiz 2153-4.

143. Janczewska I, Kapraali M, Saboonchi F, et al. Clinical application of the multigene analysis test in discriminating between ulcerative colitis and crohn's disease: A retrospective study. *Scand J Gastroenterol*. 2012;47(2):162-169.

144. Lee JC, Lyons PA, McKinney EF, et al. Gene expression profiling of CD8+ T cells predicts prognosis in patients with crohn disease and ulcerative colitis. *J Clin Invest*. 2011;121(10):4170-4179.

145. Lee JC, Espeli M, Anderson CA, et al. Human SNP links differential outcomes in inflammatory and infectious disease to a FOXO3-regulated pathway. *Cell*. 2013;155(1):57-69.

146. Ching T, Huang S, Garmire LX. Power analysis and sample size estimation for RNA-seq differential expression. *RNA*. 2014;20(11):1684-1696.

147. Yu L, Fernandez S, Brock G. Power analysis for RNA-seq differential expression studies. *BMC Bioinformatics*. 2017;18(1):234-017-1648-2.

148. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*. 2005;102(43):15545-15550.

149. Mokry M, Middendorp S, Wiegerinck CL, et al. Many inflammatory bowel disease risk loci include regions that regulate gene expression in immune cells and the intestinal epithelium. *Gastroenterology*. 2014;146(4):1040-1047.

150. Jostins L, Ripke S, Weersma RK, et al. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature*. 2012;491(7422):119-124.

151. Noble CL, Abbas AR, Lees CW, et al. Characterization of intestinal gene expression profiles in crohn's disease by genome-wide microarray analysis. *Inflamm Bowel Dis*. 2010;16(10):1717-1728.

152. Kotka M, Lieden A, Pettersson S, Trinchieri V, Masci A, D'Amato M. Solute carriers (SLC) in inflammatory bowel disease: A potential target of probiotics? *J Clin Gastroenterol*. 2008;42 Suppl 3 Pt 1:S133-5.

153. Wojtal KA, Eloranta JJ, Hruz P, et al. Changes in mRNA expression levels of solute carrier transporters in inflammatory bowel disease patients. *Drug Metab Dispos*. 2009;37(9):1871-1877.

154. Rao R, Samak G. Role of glutamine in protection of intestinal epithelial tight junctions. *J Epithel Biol Pharmacol*. 2012;5(Suppl 1-M7):47-54.

155. Dignass AU. Mechanisms and modulation of intestinal epithelial repair. *Inflamm Bowel Dis*. 2001;7(1):68-77.

156. Li X, Commane M, Nie H, et al. Act1, an NF-kappa B-activating protein. *Proc Natl Acad Sci U S A*. 2000;97(19):10489-10493.

157. Yang CW, Hojer CD, Zhou M, et al. Regulation of T cell receptor signaling by DENND1B in TH2 cells and allergic disease. *Cell*. 2016;164(1-2):141-155.

158. Shui JW, Steinberg MW, Kronenberg M. Regulation of inflammation, autoimmunity, and infection immunity by HVEM-BTLA signaling. *J Leukoc Biol*. 2011;89(4):517-523.

159. Shui JW, Kronenberg M. HVEM is a TNF receptor with multiple regulatory roles in the mucosal immune system. *Immune Netw*. 2014;14(2):67-72.

160. Tomar A, George S, Kansal P, Wang Y, Khurana S. Interaction of phospholipase C-gamma1 with villin regulates epithelial cell migration. *J Biol Chem*. 2006;281(42):31972-31986.

161. Wang Y, Srinivasan K, Siddiqui MR, George SP, Tomar A, Khurana S. A novel role for villin in intestinal epithelial cell survival and homeostasis. *J Biol Chem*. 2008;283(14):9454-9464.

162. Wang Y, George SP, Roy S, Pham E, Esmaeilniakooshkghazi A, Khurana S. Both the anti- and pro-apoptotic functions of villin regulate cell turnover and intestinal homeostasis. *Sci Rep*. 2016;6:35491.

163. Tao M, Scacheri PC, Marinis JM, Harhaj EW, Matesic LE, Abbott DW. ITCH K63-ubiquitinates the NOD2 binding protein, RIP2, to influence inflammatory signaling pathways. *Curr Biol*. 2009;19(15):1255-1263.

164. Caughey GH. Mast cell tryptases and chymases in inflammation and host defense. *Immunol Rev*. 2007;217:141-154.

165. Slattery ML, Pellatt DF, Mullany LE, Wolff RK. Differential gene expression in colon tissue associated with diet, lifestyle, and related oxidative stress. *PLoS One*. 2015;10(7):e0134406.

166. Knight JM, Kim E, Ivanov I, et al. Comprehensive site-specific whole genome profiling of stromal and epithelial colonic gene signatures in human sigmoid colon and rectal tissue. *Physiol Genomics*. 2016;48(9):651-659.

167. Mirza AH, Berthelsen CH, Seemann SE, et al. Transcriptomic landscape of lncRNAs in inflammatory bowel disease. *Genome Med*. 2015;7(1):39-015-0162-2. eCollection 2015.

168. Franke A, McGovern DP, Barrett JC, et al. Genome-wide meta-analysis increases to 71 the number of confirmed crohn's disease susceptibility loci. *Nat Genet*. 2010;42(12):1118-1125.

169. Sleiman PM, Flory J, Imielinski M, et al. Variants of DENND1B associated with asthma in children. *N Engl J Med*. 2010;362(1):36-44.

170. Smith NL, Hankinson J, Simpson A, Denning DW, Bowyer P. Reduced expression of TLR3, TLR10 and TREM1 by human macrophages in chronic cavitary pulmonary aspergillosis, and novel associations of VEGFA, DENND1B and PLAT. *Clin Microbiol Infect*. 2014;20(11):O960-8.

171. Sedy JR, Gavrieli M, Potter KG, et al. B and T lymphocyte attenuator regulates T cell activation through interaction with herpesvirus entry mediator. *Nat Immunol*. 2005;6(1):90-98.

172. Cohavy O, Zhou J, Granger SW, Ware CF, Targan SR. LIGHT expression by mucosal T cells may regulate IFN-gamma expression in the intestine. *J Immunol*. 2004;173(1):251-258.

173. Cheung TC, Humphreys IR, Potter KG, et al. Evolutionarily divergent herpesviruses modulate T cell activation by targeting the herpesvirus entry mediator cosignaling pathway. *Proc Natl Acad Sci U S A*. 2005;102(37):13218-13223.

174. Mauri DN, Ebner R, Montgomery RI, et al. LIGHT, a new member of the TNF superfamily, and lymphotoxin alpha are ligands for herpesvirus entry mediator. *Immunity*. 1998;8(1):21-30.

175. Cai G, Anumanthan A, Brown JA, Greenfield EA, Zhu B, Freeman GJ. CD160 inhibits activation of human CD4+ T cells through interaction with herpesvirus entry mediator. *Nat Immunol*. 2008;9(2):176-185.

176. Newsholme P, Curi R, Pithon Curi TC, Murphy CJ, Garcia C, Pires de Melo M. Glutamine metabolism by lymphocytes, macrophages, and neutrophils: Its importance in health and disease. *J Nutr Biochem*. 1999;10(6):316-324.

177. Carr EL, Kelman A, Wu GS, et al. Glutamine uptake and metabolism are coordinately regulated by ERK/MAPK during T lymphocyte activation. *J Immunol*. 2010;185(2):1037-1044.

178. Hume DA, Weidemann MJ. Role and regulation of glucose metabolism in proliferating cells. *J Natl Cancer Inst*. 1979;62(1):3-8.

179. Coeffier M, Miralles-Barrachina O, Le Pessot F, et al. Influence of glutamine on cytokine production by human gut in vitro. *Cytokine*. 2001;13(3):148-154.

180. Spittler A, Winkler S, Gotzinger P, et al. Influence of glutamine on the phenotype and function of human monocytes. *Blood*. 1995;86(4):1564-1569.

181. dos Santos R, Viana ML, Generoso SV, Arantes RE, Davisson Correia MI, Cardoso VN. Glutamine supplementation decreases intestinal permeability and preserves gut mucosa integrity in an experimental mouse model. *JPEN J Parenter Enteral Nutr*. 2010;34(4):408-413.

182. Seth A, Basuroy S, Sheth P, Rao RK. L-glutamine ameliorates acetaldehyde-induced increase in paracellular permeability in caco-2 cell monolayer. *Am J Physiol Gastrointest Liver Physiol*. 2004;287(3):G510-7.

183. Sido B, Seel C, Hochlehnert A, Breitkreutz R, Droge W. Low intestinal glutamine level and low glutaminase activity in crohn's disease: A rational for glutamine supplementation? *Dig Dis Sci*. 2006;51(12):2170-2179.

184. Negroni A, Cucchiara S, Stronati L. Apoptosis, necrosis, and necroptosis in the gut and intestinal homeostasis. *Mediators Inflamm*. 2015;2015:250762.

185. Nieuw Amerongen AV, Bolscher JG, Bloemena E, Veerman EC. Sulfomucins in the human body. *Biol Chem*. 1998;379(1):1-18.

186. Kindon H, Pothoulakis C, Thim L, Lynch-Devaney K, Podolsky DK. Trefoil peptide protection of intestinal epithelial barrier function: Cooperative interaction with mucin glycoprotein. *Gastroenterology*. 1995;109(2):516-523.

187. Corfield AP, Myerscough N, Bradfield N, et al. Colonic mucins in ulcerative colitis: Evidence for loss of sulfation. *Glycoconj J*. 1996;13(5):809-822.

188. Croix JA, Bhatia S, Gaskins HR. Inflammatory cues modulate the expression of secretory product genes, golgi sulfotransferases and sulfomucin production in LS174T cells. *Exp Biol Med (Maywood)*. 2011;236(12):1402-1412.

189. Godec J, Tan Y, Liberzon A, et al. Compendium of immune signatures identifies conserved and species-specific biology in response to inflammation. *Immunity*. 2016;44(1):194-206.

190. Rhein P, Scheid S, Ratei R, et al. Gene expression shift towards normal B cells, decreased proliferative capacity and distinct surface receptors characterize leukemic blasts persisting during induction therapy in childhood acute lymphoblastic leukemia. *Leukemia*. 2007;21(5):897-905.

191. van der Hulst RR, van Kreel BK, von Meyenfeldt MF, et al. Glutamine and the preservation of gut integrity. *Lancet*. 1993;341(8857):1363-1365.

192. Laura Irvin, Roschelle Heuberger. Enhancing gut function and providing symptom relief in IBD with glutamine supplementation: a literature review. *Gastrointestinal Nursing* 2015;13(6).

193. Sesto A, Navarro M, Burslem F, Jorcano JL. Analysis of the ultraviolet B response in primary human keratinocytes using oligonucleotide microarrays. *Proc Natl Acad Sci U S A*. 2002;99(5):2965-2970.

194. Ulmer AJ, Flad H, Rietschel T, Mattern T. Induction of proliferation and cytokine production in human T lymphocytes by lipopolysaccharide (LPS). *Toxicology*. 2000;152(1-3):37-45.

195. Obata Y, Takahashi D, Ebisawa M, et al. Epithelial cell-intrinsic notch signaling plays an essential role in the maintenance of gut immune homeostasis. *J Immunol*. 2012;188(5):2427-2436.

196. Zheng X, Tsuchiya K, Okamoto R, et al. Suppression of hath1 gene expression directly regulated by hes1 via notch signaling is associated with

goblet cell depletion in ulcerative colitis. *Inflamm Bowel Dis*. 2011;17(11):2251-2260.

197. Danielson PB. The cytochrome P450 superfamily: Biochemistry, evolution and drug metabolism in humans. *Curr Drug Metab*. 2002;3(6):561-597.

198. Fujiwara R, Yokoi T, Nakajima M. Structure and protein-protein interactions of human UDP-glucuronosyltransferases. *Front Pharmacol*. 2016;7:388.

199. Luo Y, de Lange KM, Jostins L, et al. Exploring the genetic architecture of inflammatory bowel disease by whole-genome sequencing identifies association at ADCY7. *Nat Genet*. 2017.

200. Neer EJ. Heterotrimeric G proteins: Organizers of transmembrane signals. *Cell*. 1995;80(2):249-257.

201. Denker BM, Saha C, Khawaja S, Nigam SK. Involvement of a heterotrimeric G protein alpha subunit in tight junction biogenesis. *J Biol Chem*. 1996;271(42):25750-25753.

202. Gibbons DL, Abeler-Dorner L, Raine T, et al. Cutting edge: Regulator of G protein signaling-1 selectively regulates gut T cell trafficking and colitic potential. *J Immunol*. 2011;187(5):2067-2071.

203. Sabath E, Negoro H, Beaudry S, et al. Galpha12 regulates protein interactions within the MDCK cell tight junction and inhibits tight-junction assembly. *J Cell Sci*. 2008;121(Pt 6):814-824.

204. Fre S, Huyghe M, Mourikis P, Robine S, Louvard D, Artavanis-Tsakonas S. Notch signals control the fate of immature progenitor cells in the intestine. *Nature*. 2005;435(7044):964-968.

205. van Es JH, van Gijn ME, Riccio O, et al. Notch/gamma-secretase inhibition turns proliferative cells in intestinal crypts and adenomas into goblet cells. *Nature*. 2005;435(7044):959-963.

206. Pellegrinet L, Rodilla V, Liu Z, et al. Dll1- and dll4-mediated notch signaling are required for homeostasis of intestinal stem cells. *Gastroenterology*. 2011;140(4):1230-1240.e1-7.

207. VanDussen KL, Carulli AJ, Keeley TM, et al. Notch signaling modulates proliferation and differentiation of intestinal crypt base columnar stem cells. *Development*. 2012;139(3):488-497.

208. Xing Y, Chen X, Cao Y, Huang J, Xie X, Wei Y. Expression of wnt and notch signaling pathways in inflammatory bowel disease treated with mesenchymal stem cell transplantation: Evaluation in a rat model. *Stem Cell Res Ther*. 2015;6:101-015-0092-3.

209. Sun Y, Lowther W, Kato K, et al. Notch4 intracellular domain binding to Smad3 and inhibition of the TGF-beta signaling. *Oncogene*. 2005;24(34):5365-5374.

210. Mathern DR, Laitman LE, Hovhannisyan Z, et al. Mouse and human notch-1 regulate mucosal immune responses. *Mucosal Immunol*. 2014;7(4):995-1005.

211. Rusanescu G, Mao J. Notch3 is necessary for neuronal differentiation and maturation in the adult spinal cord. *J Cell Mol Med*. 2014;18(10):2103-2116.

212. Dang TP, Eichenberger S, Gonzalez A, Olson S, Carbone DP. Constitutive activation of Notch3 inhibits terminal epithelial differentiation in lungs of transgenic mice. *Oncogene*. 2003;22(13):1988-1997.

213. Anastasi E, Campese AF, Bellavia D, et al. Expression of activated Notch3 in transgenic mice enhances generation of T regulatory cells and protects

against experimental autoimmune diabetes. *J Immunol*. 2003;171(9):4504-4511.

214. Green JT, Rhodes J, Ragunath K, et al. Clinical status of ulcerative colitis in patients who smoke. *Am J Gastroenterol*. 1998;93(9):1463-1467.

215. Cottone M, Rosselli M, Orlando A, et al. Smoking habits and recurrence in crohn's disease. *Gastroenterology*. 1994;106(3):643-648.

216. Biedermann L, Fournier N, Misselwitz B, et al. High rates of smoking especially in female crohn's disease patients and low use of supportive measures to achieve smoking cessation--data from the swiss IBD cohort study. *J Crohns Colitis*. 2015;9(10):819-829.

217. McCrea KA, Ensor JE, Nall K, Bleecker ER, Hasday JD. Altered cytokine regulation in the lungs of cigarette smokers. *Am J Respir Crit Care Med*. 1994;150(3):696-703.

218. Miller LG, Goldstein G, Murphy M, Ginns LC. Reversible alterations in immunoregulatory T cells in smoking. analysis by monoclonal antibodies and flow cytometry. *Chest*. 1982;82(5):526-529.

219. Sopori M. Effects of cigarette smoke on the immune system. *Nat Rev Immunol*. 2002;2(5):372-377.

220. Green JT, McKirdy HC, Rhodes J, Thomas GA, Evans BK. Intra-luminal nicotine reduces smooth muscle tone and contractile activity in the distal large bowel. *Eur J Gastroenterol Hepatol*. 1999;11(11):1299-1304.

221. Srivastava ED, Russell MA, Feyerabend C, Rhodes J. Effect of ulcerative colitis and smoking on rectal blood flow. *Gut*. 1990;31(9):1021-1024.

222. Machiela MJ, Chanock SJ. LDlink: A web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. *Bioinformatics*. 2015;31(21):3555-3557.

223. Wu S, Trievel RC, Rice JC. Human SFMBT is a transcriptional repressor protein that selectively binds the N-terminal tail of histone H3. *FEBS Lett*. 2007;581(17):3289-3296.

224. De Vries L, Zheng B, Fischer T, Elenko E, Farquhar MG. The regulator of G protein signaling family. *Annu Rev Pharmacol Toxicol*. 2000;40:235-271.

225. Xie X, Wang Z, Chen Y. Association of LKB1 with a WD-repeat protein WDR6 is implicated in cell growth arrest and p27(Kip1) induction. *Mol Cell Biochem*. 2007;301(1-2):115-122.

226. Islam MM, Suzuki H, Yoneda M, Tanaka M. Primary structure of the smallest (6.4-kDa) subunit of human and bovine ubiquinol-cytochrome c reductase deduced from cDNA sequences. *Biochem Mol Biol Int*. 1997;41(6):1109-1116.

227. Petroziello J, Yamane A, Westendorf L, et al. Suppression subtractive hybridization and expression profiling identifies a unique set of genes overexpressed in non-small-cell lung cancer. *Oncogene*. 2004;23(46):7734-7745.

228. Gilli F, Lindberg RL, Valentino P, et al. Learning from nature: Pregnancy changes the expression of inflammation-related genes in patients with multiple sclerosis. *PLoS One*. 2010;5(1):e8962.

229. Nagarajan NA, Gonzalez F, Shastri N. Nonclassical MHC class ib-restricted cytotoxic T cells monitor antigen processing in the endoplasmic reticulum. *Nat Immunol*. 2012;13(6):579-586.

230. Gomes LC, Scorrano L. High levels of Fis1, a pro-fission mitochondrial protein, trigger autophagy. *Biochim Biophys Acta*. 2008;1777(7-8):860-866.

231. Li L, Bin LH, Li F, et al. TRIP6 is a RIP2-associated common signaling component of multiple NF-kappaB activation pathways. *J Cell Sci*. 2005;118(Pt 3):555-563.

232. Vidal-Taboada JM, Lu A, Pique M, Pons G, Gil J, Oliva R. Down syndrome critical region gene 2: Expression during mouse development and in human cell lines indicates a function related to cell proliferation. *Biochem Biophys Res Commun*. 2000;272(1):156-163.

233. Xu J, Lai YJ, Lin WC, Lin FT. TRIP6 enhances lysophosphatidic acid-induced cell migration by interacting with the lysophosphatidic acid 2 receptor. *J Biol Chem*. 2004;279(11):10459-10468.

234. Trusolino L, Bertotti A, Comoglio PM. MET signalling: Principles and functions in development, organ regeneration and cancer. *Nat Rev Mol Cell Biol*. 2010;11(12):834-848.

235. Lee S, Park YY, Kim SH, et al. Human mitochondrial Fis1 links to cell cycle regulators at G2/M transition. *Cell Mol Life Sci*. 2014;71(4):711-725.

236. Wu S, Trievel RC, Rice JC. Human SFMBT is a transcriptional repressor protein that selectively binds the N-terminal tail of histone H3. *FEBS Lett*. 2007;581(17):3289-3296.

237. Chang YT, Chou CT, Shiao YM, et al. Psoriasis vulgaris in chinese individuals is associated with PSORS1C3 and CDSN genes. *Br J Dermatol*. 2006;155(4):663-669.

238. Zhang Y, Xu YZ, Sun N, et al. Long noncoding RNA expression profile in fibroblast-like synoviocytes from patients with rheumatoid arthritis. *Arthritis Res Ther*. 2016;18(1):227.

239. Wang CY, Liang YJ, Lin YS, Shih HM, Jou YS, Yu WC. YY1AP, a novel co-activator of YY1. *J Biol Chem*. 2004;279(17):17750-17755.

240. Salomon B, Lenschow DJ, Rhee L, et al. B7/CD28 costimulation is essential for the homeostasis of the CD4+CD25+ immunoregulatory T cells that control autoimmune diabetes. *Immunity*. 2000;12(4):431-440.

241. Xie X, Wang Z, Chen Y. Association of LKB1 with a WD-repeat protein WDR6 is implicated in cell growth arrest and p27(Kip1) induction. *Mol Cell Biochem*. 2007;301(1-2):115-122.

242. McKnight NC, Jefferies HB, Alemu EA, et al. Genome-wide siRNA screen reveals amino acid starvation-induced autophagy requires SCOC and WAC. *EMBO J*. 2012;31(8):1931-1946.

243. Roy PK, Rashid F, Bragg J, Ibdah JA. Role of the JNK signal transduction pathway in inflammatory bowel disease. *World J Gastroenterol*. 2008;14(2):200-202.

244. Schroeder HW,Jr, Cavacini L. Structure and function of immunoglobulins. *J Allergy Clin Immunol*. 2010;125(2 Suppl 2):S41-52.

245. Kochan G, Krojer T, Harvey D, et al. Crystal structures of the endoplasmic reticulum aminopeptidase-1 (ERAP1) reveal the molecular basis for N-terminal peptide trimming. *Proc Natl Acad Sci U S A*. 2011;108(19):7745-7750.

246. Birtley JR, Saridakis E, Stratikos E, Mavridis IM. The crystal structure of human endoplasmic reticulum aminopeptidase 2 reveals the atomic basis for distinct roles in antigen processing. *Biochemistry*. 2012;51(1):286-295.

247. Genetic Analysis of Psoriasis Consortium & the Wellcome Trust Case Control Consortium 2, Strange A, Capon F, et al. A genome-wide association study identifies new psoriasis susceptibility loci and an interaction between HLA-C and ERAP1. *Nat Genet*. 2010;42(11):985-990.

248. Kirino Y, Bertsias G, Ishigatsubo Y, et al. Genome-wide association analysis identifies new susceptibility loci for behcet's disease and epistasis between HLA-B*51 and ERAP1. *Nat Genet*. 2013;45(2):202-207.

249. Franke A, McGovern DP, Barrett JC, et al. Genome-wide meta-analysis increases to 71 the number of confirmed crohn's disease susceptibility loci. *Nat Genet*. 2010;42(12):1118-1125.

250. Saeki N, Kim DH, Usui T, et al. GASDERMIN, suppressed frequently in gastric cancer, is a target of LMO1 in TGF-beta-dependent apoptotic signalling. *Oncogene*. 2007;26(45):6488-6498.

251. Chao KL, Kulakova L, Herzberg O. Gene polymorphism linked to increased asthma and IBD risk alters gasdermin-B structure, a sulfatide and phosphoinositide binding protein. *Proc Natl Acad Sci U S A*. 2017;114(7):E1128-E1137.

252. Martinelli N, Girelli D, Malerba G, et al. FADS genotypes and desaturase activity estimated by the ratio of arachidonic acid to linoleic acid are associated with inflammation and coronary artery disease. *Am J Clin Nutr*. 2008;88(4):941-949.

253. Peters JE, Lyons PA, Lee JC, et al. Insight into genotype-phenotype associations through eQTL mapping in multiple cell types in health and immune-mediated disease. *PLoS Genet*. 2016;12(3):e1005908.

254. Stroud CK, Nara TY, Roqueta-Rivera M, et al. Disruption of FADS2 gene in mice impairs male reproduction and causes dermal and intestinal ulceration. *J Lipid Res*. 2009;50(9):1870-1880.

255. Pu J, Schindler C, Jia R, Jarnik M, Backlund P, Bonifacino JS. BORC, a multisubunit complex that regulates lysosome positioning. *Dev Cell*. 2015;33(2):176-188.

256. Li Y, Xie X. A mixture model for expression deconvolution from RNA-seq in heterogeneous tissues. *BMC Bioinformatics*. 2013;14 Suppl 5:S11-2105-14-S5-S11. Epub 2013 Apr 10.

257. Steuerman Y, Gat-Viks I. Exploiting gene-expression deconvolution to probe the genetics of the immune system. *PLoS Comput Biol*. 2016;12(4):e1004856.

258. Wen Y, Wei Y, Zhang S, et al. Cell subpopulation deconvolution reveals breast cancer heterogeneity based on DNA methylation signature. *Brief Bioinform*. 2016.

259. Koestler DC, Jones MJ, Usset J, et al. Improving cell mixture deconvolution by identifying optimal DNA methylation libraries (IDOL). *BMC Bioinformatics*. 2016;17:120-016-0943-7.

260. Venet D, Pecasse F, Maenhaut C, Bersini H. Separation of samples into their constituents using gene expression data. *Bioinformatics*. 2001;17 Suppl 1:S279-87.

261. Abbas AR, Wolslegel K, Seshasayee D, Modrusan Z, Clark HF. Deconvolution of blood microarray data identifies cellular activation patterns in systemic lupus erythematosus. *PLoS One*. 2009;4(7):e6098.

262. Shen-Orr SS, Tibshirani R, Khatri P, et al. Cell type-specific gene expression differences in complex tissues. *Nat Methods*. 2010;7(4):287-289.

263. Repsilber D, Kern S, Telaar A, et al. Biomarker discovery in heterogeneous tissue samples -taking the in-silico deconfounding approach. *BMC Bioinformatics*. 2010;11:27-2105-11-27.

264. Wong L, Jiang K, Chen Y, et al. Limits of peripheral blood mononuclear cells for gene expression-based biomarkers in juvenile idiopathic arthritis. *Sci Rep*. 2016;6:29477.

265. Larmonier CB, Shehab KW, Ghishan FK, Kiela PR. T lymphocyte dynamics in inflammatory bowel diseases: Role of the microbiome. *Biomed Res Int*. 2015;2015:504638.

266. Cheroutre H. In IBD eight can come before four. *Gastroenterology*. 2006;131(2):667-670.

267. Mahida YR. The key role of macrophages in the immunopathogenesis of inflammatory bowel disease. *Inflamm Bowel Dis*. 2000;6(1):21-33.

268. Pauley KM, Cha S, Chan EK. MicroRNA in autoimmunity and autoimmune diseases. *J Autoimmun*. 2009;32(3-4):189-194.

Appendix 1 – Principle component analysis script

```
#Load the appropriate libraries
library("edgeR")
library("ggplot2")
library("reshape2")
library("Cairo")

#Import the phenotype data
metadata=
read.delim2("/home/demandtl/RNAseqLD/DE/phenotypedata_20160530.csv",heade
r=TRUE,colClasses="character",sep=",")

# Add a 'class' column with 'healthy' or 'ibd'
metadata = metadata[order(metadata$type),]
metadata$class = "healthy"
wIBD = which(metadata$type!="control")
metadata[wIBD,"class"]="ibd"

#Convert date of birth (DOB) to age
Dob = as.Date(metadata$dob, format ="%m/%d/%Y")
age = (as.Date("05/04/2016", format ="%m/%d/%Y") - Dob)/365
metadata$age = round(age)

#Save phenotype data after changes
write.csv(x=metadata,row.names = TRUE,
file=paste("/home/demandtl/RNAseqLD/DE/PCA plots 128
libraries/correctedMetadata.csv", sep=""))

# Import the gene expression count table
sampleFiles=sort(list.files(path="/home/demandtl/RNAseqLD/DE/allHTSeq/"))
countsOriginal=readDGE(files=sampleFiles,path="/home/demandtl/RNAseqLD/DE
/allHTSeq/")$counts

#Check files loaded correctly
head(countsOriginal)
tail(countsOriginal)

#Save the counts table
write.csv(x=countsOriginal,row.names = TRUE,
file=paste("/home/demandtl/RNAseqLD/DE/PCA plots 128
libraries/countsOriginal.csv", sep=""))

#Convert counts to counts per-million
cpmsOriginal = cpm(countsOriginal );

#Remove all 'unwanted' rows and set cut-off for gene expression levels
noint = rownames(countsOriginal) %in%
c("no_feature","ambiguous","too_low_aQual","not_aligned","alignment_not_unique")

keep = rowSums(cpmsOriginal > 1) >= 24 & !noint
```

```
counts = countsOriginal[keep,]
cpms = cpmsOriginal[keep,]

#Check file was edited correctly
dim(cpms)
head(counts)
tail(counts)

#Save the new expression counts table
write.csv(x=counts,row.names = TRUE,
file=paste("/home/demandtl/RNAseqLD/DE/PCA plots 128
libraries/edgeRFilteredCountsTable.csv", sep=""))

#Calculate PCs with normalized counts
d = DGEList(counts=counts , group=metadata$class )
d = calcNormFactors(d)
nc = cpm(d, normalized.lib.sizes=TRUE)
gNames = rownames(nc)

pca<-prcomp(t(nc),center=TRUE, scale=FALSE)

#Check that the order of the PC data is the same of the metadata table
rownames(pca$x)==metadata$sampleName

#Create dataframe with PCs and relevant phenotype data
dfPCA = data.frame(sampleName=rownames(pca$x), PC1=pca$x[,1],
PC2=pca$x[,2], PC3=pca$x[,3], PC4=pca$x[,4], PC5=pca$x[,5], PC6=pca$x[,6])

dfPcaAndMetadata = merge(x=dfPCA, y=metadata, by.x=0 , by.y="sampleName")

#PC1 vs PC2 for batch, sex and disease type
a = ggplot(data=dfPcaAndMetadata, aes(x=PC1, y=PC2, color=batch)) +
geom_point(shape=19, size=3) + stat_ellipse(type="norm", level=0.68)

b = ggplot(data=dfPcaAndMetadata, aes(x=PC1, y=PC2, color=sex)) +
geom_point(shape=19, size=3) + stat_ellipse(type="norm", level=0.68)

c = ggplot(data=dfPcaAndMetadata, aes(x=PC1, y=PC2, color=type)) +
geom_point(shape=19, size=3) + stat_ellipse(type="norm", level=0.68)

#PC3 vs PC4 for batch, sex and disease type
d = ggplot(data=dfPcaAndMetadata, aes(x=PC3, y=PC4, color=batch)) +
geom_point(shape=19, size=3) + stat_ellipse(type="norm", level=0.68)

e = ggplot(data=dfPcaAndMetadata, aes(x=PC3, y=PC4, color=sex)) +
geom_point(shape=19, size=3) + stat_ellipse(type="norm", level=0.68)

f = ggplot(data=dfPcaAndMetadata, aes(x=PC3, y=PC4, color=type)) +
geom_point(shape=19, size=3) + stat_ellipse(type="norm", level=0.68)

#PC5 vs PC6 for batch, sex and disease type
```

```
g = ggplot(data=dfPcaAndMetadata, aes(x=PC5, y=PC6, color=batch)) +
geom_point(shape=19, size=3) + stat_ellipse(type="norm", level=0.68)

h = ggplot(data=dfPcaAndMetadata, aes(x=PC5, y=PC6, color=sex)) +
geom_point(shape=19, size=3) + stat_ellipse(type="norm", level=0.68)

i = ggplot(data=dfPcaAndMetadata, aes(x=PC5, y=PC6, color=type)) +
geom_point(shape=19, size=3) + stat_ellipse(type="norm", level=0.68)

#Visualise and save the plots (perform for plots a-i)
print(a)

ggsave(plot=a, filename = paste("/home/demandtl/RNAseqLD/DE/PCA plots 128
libraries/PCAplot_1_2_batch.png", sep=""), height=10, width=10*1.2, dpi=222)
```

Appendix 2 – Differential expression analysis script

```
#Load the appropriate libraries
library("ShortRead")
library("DESeq")
library("edgeR")
library("ggplot2")
library("reshape2")
library("Cairo")
library("RUVSeq")
library("GGally")
library("biomaRt")


#Load the samples and check the number of samples present
sampleFiles=sort(list.files(path="/home/demandtl/RNAseqLD/DE/allHTSeq/"))
nSamples=length(sampleFiles)

#Import the formatted phenotype data
metadata=
read.delim2("/home/demandtl/RNAseqLD/DE/correctedMetadata.csv",header=TRU
E,colClasses="character",sep=",")

#Remove the sample identified as an outlier in the PC analysis
w1740 = which(metadata$sampleName == "GKT1740")
metadata=metadata[-w1740, ]

#Save phenotype data after removal outlier
runDate = paste(unlist(strsplit(as.character(Sys.Date()), split = "-")), collapse="")
write.csv(x=metadata,row.names = TRUE,
file=paste("/home/demandtl/RNAseqLD/DE/correctedMetadata_",runDate,".csv",
sep=""))

# Import the gene expression count table
sampleFiles=sort(list.files(path="/home/demandtl/RNAseqLD/DE/allHTSeq/"))
countsOriginal=readDGE(files=sampleFiles,path="/home/demandtl/RNAseqLD/DE
/allHTSeq/")$counts

#Check files loaded correctly
head(countsOriginal)
tail(countsOriginal)

#Save the counts table
write.csv(x=countsOriginal,row.names = TRUE,
file=paste("/home/demandtl/RNAseqLD/DE/PCA plots 128
libraries/countsOriginal.csv", sep=""))

#Convert counts to counts per-million
cpmsOriginal = cpm(countsOriginal );

#Remove all 'unwanted' rows and set cut-off for gene expression levels
```

```
noint = rownames(countsOriginal) %in%
c("no_feature","ambiguous","too_low_aQual","not_aligned","alignment_not_unique")

keep = rowSums(cpmsOriginal > 1) >= 24 & !noint
counts = countsOriginal[keep,]
cpms = cpmsOriginal[keep,]

#Check file was edited correctly
dim(cpms)
head(counts)
tail(counts)

#Save the new expression counts table
write.csv(x=counts,row.names = TRUE,
file=paste("/home/demandtl/RNAseqLD/DE/PCA plots 128
libraries/edgeRFilteredCountsTable.csv", sep=""))

#Calculate RUV values
geneEnsgIds = rownames(counts)[grep("^ENSG", rownames(counts))]
spikeIds =rownames(counts)[grep("^ERCC", rownames(counts))]
ruv = RUVg(x=counts, cIdx=spikeIds, k=6)
counts2 = ruv$normalizedCounts

metadata$w1 = ruv$W[,1]
metadata$w2 = ruv$W[,2]
metadata$w3 = ruv$W[,3]
metadata$w4 = ruv$W[,4]
metadata$w5 = ruv$W[,5]
metadata$w6 = ruv$W[,6]

#Correct count data using RUVseq and PCA values
d = DGEList(counts=counts, group=metadata$class)
d = calcNormFactors(d, method="TMM" )
design = model.matrix(~ w1 + w2 + w3 + w4 + w5 + w6 + batch + sex + age +
class, data=metadata)
d = estimateGLMRobustDisp(d, design)
summary(d$tagwise.dispersion)
f = glmFit(d, design)
IBD_ruvBatch = glmLRT(f, coef="classibd ")

#Correct for multiple testing
resultsTable = IBD_ruvBatch$table
q = p.adjust(resultsTable$PValue, method="BH")
resultsTable$QValue = q

#Import gene names from Ensemble into results table
listMarts(host = "www.ensembl.org")
ensembl =
useMart("ENSEMBL_MART_ENSEMBL",dataset="hsapiens_gene_ensembl", host =
"www.ensembl.org")
ensgIds = rownames(resultsTable)
```

```
geneSymbol = getBM(attributes=c("hgnc_symbol", "ensembl_gene_id"),
filters="ensembl_gene_id", values=ensgIds, mart=ensembl)

resultsTable$geneSymbol = "NA"

for(i in 1:nrow(geneSymbol)){gS = geneSymbol[i,"hgnc_symbol"]
gE = geneSymbol[i,"ensembl_gene_id"]
w= which(rownames(resultsTable) == gE)
resultsTable[w, "geneSymbol"]=gS
print(i)
}

w = which(colnames(resultsTable)=="LR")
resultsTable = resultsTable[,-w]

#Generate and save the volcano plot
ggp = ggplot(data=resultsTable) + geom_hline(mapping = aes(yintercept=-
log10(0.05)), col="darkgreen", linetype = 2) + geom_point(aes(x=logFC, y=I(-
log10(QValue)), col=logCPM ), alpha=0.8, size=2, pch=19) + ylab("-log10(q-
value)") + xlab("log(Fold Change)") + annotate("text", label = "q-value = 0.05", x
= -2.3, y = 1.45, size = 4, colour = "darkgreen")

print(ggp)

ggsave(filename =
paste("/home/demandtl/RNAseqLD/DE/volcanoPlot_",runDate,".png",sep=""),
height=8, width=8, dpi=222, plot=ggp)

#Order the results table by significance (q-value)
o = order(resultsTable$QValue)
resultsTable = resultsTable[o, ]

#Check ordering worked and save/export results table
head(resultsTable)

write.csv(x=resultsTable, row.names = TRUE, quote=FALSE,
file=paste("/home/demandtl/RNAseqLD/DE/Diffexp_IBDvsCont_",runDate,".csv",
sep=""))
```

Appendix 3 – Matrix eQTL script

```
#Load appropriate libraries
library("MatrixEQTL")
setwd("/gpfs/home/demandtl/RNAseqLD/eQTL/")
base.dir = "/gpfs/home/demandtl/RNAseqLD/eQTL/"

useModel = modelLINEAR

#Set path to genotype files and SNP location file
SNP_file_name = paste(base.dir, "genotypes_eqtl_filtered60_output.txt", sep="")
snps_location_file_name = paste(base.dir, "SNP locations_filtered.txt", sep="")

#Set path to gene expression file and gene location file
expression_file_name = paste(base.dir, "FPKM_aboveBackground.txt", sep="")
gene_location_file_name = paste(base.dir, "Gene locations_filtered.txt", sep="")

#Set path to covariate file (PC data)
covariates_file_name = paste(base.dir, "PCA_variables_table.txt", sep="")
errorCovariance = numeric()

#Set the dicstance for local gene-SNP pair (cis)
cisDist = 1000000

#Set the Output files
output_file_name_cis = "/gpfs/home/demandtl/RNAseqLD/eQTL/Output/Gene-
cis_results.txt"
output_file_name_tra = "/gpfs/home/demandtl/RNAseqLD/eQTL/Output/Gene-
tra_results.txt"

#Set the significane threshold
pvOutputThreshold_cis = 0.02;
pvOutputThreshold_tra = 0.00;

#Load genotype data
snps = SlicedData$new()
snps$fileDelimiter = "\t"
snps$fileOmitCharacters = "NA"
snps$fileSkipRows = 1
snps$fileSkipColumns = 1
snps$fileSliceSize = 2000
snps$LoadFile(SNP_file_name)

#Load gene expression data
gene = SlicedData$new()
gene$fileDelimiter = "\t"
gene$fileOmitCharacters = "NA"
gene$fileSkipRows = 1
gene$fileSkipColumns = 1
gene$fileSliceSize = 2000
gene$LoadFile(expression_file_name)
```

```
#Load covariates
cvrt = SlicedData$new()
cvrt$fileDelimiter = "\t"
cvrt$fileOmitCharacters = "NA"
cvrt$fileSkipRows = 1
cvrt$fileSkipColumns = 1
if(length(covariates_file_name)>0) {cvrt$LoadFile(covariates_file_name)}

#Run the analysis
snpspos = read.table(snps_location_file_name, header = TRUE, stringsAsFactors =
FALSE);
genepos = read.table(gene_location_file_name, header = TRUE, stringsAsFactors =
FALSE);


me = Matrix_eQTL_main(
 snps = snps,
 gene = gene,
 cvrt = cvrt,
 output_file_name = output_file_name_tra,
 pvOutputThreshold = pvOutputThreshold_tra,
 useModel = useModel,
 errorCovariance = errorCovariance,
 verbose = TRUE,
 output_file_name.cis = output_file_name_cis,
 pvOutputThreshold.cis = pvOutputThreshold_cis,
 snpspos = snpspos,
 genepos = genepos,
 cisDist = cisDist,
 pvalue.hist = "qqplot",
 min.pv.by.genesnp = FALSE,
 noFDRsaveMemory = FALSE)

#View the results
cat('Analysis done in: ', me$time.in.sec, ' seconds', '\n');
cat('Detected local eQTLs:', '\n');
show(me$cis$eqtls)

#Visualise the qqplot
plot(me)
```

Appendix 4 – IBD susceptibility loci Locations

Loci no.	Chromosome	Start locus (bp)	End locus (bp)	Extended start (bp)	Extended end (bp)
1.01	chr1	1194804	1346703	694804	1846703
1.02	chr1	2470681	2514575	1970681	3014575
1.03	chr1	7969507	8186232	7469507	8686232
1.04	chr1	20133810	20227723	19633810	20727723
1.05	chr1	22681214	22711473	22181214	23211473
1.06	chr1	62900811	63204364	62400811	63704364
1.07	chr1	67598347	67743552	67098347	68243552
1.08	chr1	70991829	71040166	70491829	71540166
1.09	chr1	78450517	78623626	77950517	79123626
1.10	chr1	92554283	92554283	92054283	93054283
1.11	chr1	101293753	101575205	100793753	102075205
1.12	chr1	114303808	114377568	113803808	114877568
1.13	chr1	120437718	120638604	119937718	121138604
1.14	chr1	151792984	151802356	151292984	152302356
1.15	chr1	155612197	156011444	155112197	156511444
1.16	chr1	159800000	159890000	159300000	160390000
1.17	chr1	160837622	160919496	160337622	161419496
1.18	chr1	161463601	161479745	160963601	161979745
1.19	chr1	169090748	169519049	168590748	170019049
1.20	chr1	172803959	172870991	172303959	173370991
1.21	chr1	186862512	186967702	186362512	187467702
1.22	chr1	197342380	197813558	196842380	198313558
1.23	chr1	198598663	198670555	198098663	199170555
1.24	chr1	200065713	200105746	199565713	200605746
1.25	chr1	200874229	201024059	200374229	201524059
1.26	chr1	206939904	206968955	206439904	207468955
1.27	chr1	209970000	210020000	209470000	210520000
2.01	chr2	25075281	25161265	24575281	25661265
2.02	chr2	27598097	27752871	27098097	28252871
2.03	chr2	28602911	28647084	28102911	29147084
2.04	chr2	43517088	43850357	43017088	44350357
2.05	chr2	61186829	61231014	60686829	61731014
2.06	chr2	62551472	62575443	62051472	63075443
2.07	chr2	65604914	65692016	65104914	66192016
2.08	chr2	102610642	103094213	102110642	103594213
2.09	chr2	145417530	145627269	144917530	146127269
2.10	chr2	160691494	160878364	160191494	161378364
2.11	chr2	163110536	163124051	162610536	163624051
2.12	chr2	182310000	182330000	181810000	182830000
2.13	chr2	187500000	187680000	187000000	188180000
2.14	chr2	191907655	191972789	191407655	192472789
2.15	chr2	198244598	198954831	197744598	199454831

Loci no.	Chromosome	Start locus (bp)	End locus (bp)	Extended start (bp)	Extended end (bp)
2.16	chr2	199489760	200152198	198989760	200652198
2.17	chr2	204574890	204649276	204074890	205149276
2.18	chr2	219066980	219191569	218566980	219691569
2.19	chr2	228639557	228664568	228139557	229164568
2.20	chr2	231083171	231171423	230583171	231671423
2.21	chr2	234143048	234208258	233643048	234708258
2.22	chr2	241563739	241608453	241063739	242108453
2.23	chr2	242470000	242490000	241970000	242990000
2.24	chr2	242724543	242740537	242224543	243199373
3.01	chr3	18699977	18825669	18199977	19325669
3.02	chr3	46150937	46486611	45650937	46986611
3.03	chr3	48446237	51095279	47946237	51595279
3.04	chr3	52978418	53142980	52478418	53642980
3.05	chr3	53100000	53170000	52600000	53670000
3.06	chr3	71160000	71190000	70660000	71690000
3.07	chr3	100910000	101270000	100410000	101770000
3.08	chr3	101560223	101576029	101060223	102076029
3.09	chr3	141070000	141150000	140570000	141650000
3.10	chr3	141072289	141154542	140572289	141654542
3.11	chr3	188400000	188490000	187900000	188990000
4.01	chr4	3398068	3450541	2898068	3950541
4.02	chr4	26132361	26132361	25632361	26632361
4.03	chr4	38324347	38373273	37824347	38873273
4.04	chr4	38580000	38590000	38080000	39090000
4.05	chr4	48344930	48430354	47844930	48930354
4.06	chr4	74736180	74873602	74236180	75373602
4.07	chr4	102702364	103001649	102202364	103501649
4.08	chr4	103391275	103548216	102891275	104048216
4.09	chr4	106063987	106217358	105563987	106717358
4.10	chr4	123031494	123558828	122531494	124058828
5.01	chr5	532632	685849	32632	1185849
5.02	chr5	10670274	10759514	10170274	11259514
5.03	chr5	38855122	38881538	38355122	39381538
5.04	chr5	40219972	40623346	39719972	41123346
5.05	chr5	55436851	55442249	54936851	55942249
5.06	chr5	71683885	71747448	71183885	72247448
5.07	chr5	72502029	72559339	72002029	73059339
5.08	chr5	96200770	96373750	95700770	96873750
5.09	chr5	129723552	131833599	129223552	132333599
5.10	chr5	134422204	134453814	133922204	134953814
5.11	chr5	141435466	141543989	140935466	142043989
5.12	chr5	149590000	149630000	149090000	150130000
5.13	chr5	150169843	150338714	149669843	150838714
5.14	chr5	158764177	158856513	158264177	159356513

Loci no.	Chromosome	Start locus (bp)	End locus (bp)	Extended start (bp)	Extended end (bp)
5.15	chr5	172313034	172329734	171813034	172829734
5.16	chr5	173269956	173399325	172769956	173899325
5.17	chr5	176782218	176806636	176282218	177306636
6.01	chr6	382559	403799	0	903799
6.02	chr6	3416922	3445536	2916922	3945536
6.03	chr6	14711961	14734463	14211961	15234463
6.04	chr6	19720000	19830000	19220000	20330000
6.05	chr6	20640419	20891190	20140419	21391190
6.06	chr6	21427143	21444899	20927143	21944899
6.07	chr6	31236467	31313602	30736467	31813602
6.08	chr6	32626272	32626952	32126272	33126952
6.09	chr6	42000000	42010000	41500000	42510000
6.10	chr6	90809560	91014029	90309560	91514029
6.11	chr6	106435025	106442096	105935025	106942096
6.12	chr6	111493953	111919424	110993953	112419424
6.13	chr6	127413222	127532807	126913222	128032807
6.14	chr6	128215237	128297611	127715237	128797611
6.15	chr6	137959235	138006504	137459235	138506504
6.16	chr6	143865221	143924048	143365221	144424048
6.17	chr6	149558895	149610339	149058895	150110339
6.18	chr6	159489791	159515309	158989791	160015309
6.19	chr6	167360389	167485800	166860389	167985800
7.01	chr7	2752152	2912928	2252152	3412928
7.02	chr7	6500000	6550000	6000000	7050000
7.03	chr7	17430004	17445706	16930004	17945706
7.04	chr7	20580000	20589000	20080000	21089000
7.05	chr7	26694926	26911904	26194926	27411904
7.06	chr7	27231762	27248891	26731762	27748891
7.07	chr7	28142088	28214300	27642088	28714300
7.08	chr7	50096251	50323456	49596251	50823456
7.09	chr7	98724730	98785080	98224730	99285080
7.1	chr7	100401433	100433794	99901433	100933794
7.11	chr7	107437613	107584780	106937613	108084780
7.12	chr7	116889718	116917118	116389718	117417118
7.13	chr7	128567032	128581835	128067032	129081835
7.14	chr7	148211140	148251668	147711140	148751668
7.15	chr7	148400000	148580000	147900000	149080000
8.01	chr8	27189213	27303015	26689213	27803015
8.02	chr8	49047317	49206630	48547317	49706630
8.03	chr8	90854846	90877546	90354846	91377546
8.04	chr8	126529074	126541090	126029074	127041090
8.05	chr8	129501028	129571140	129001028	130071140
8.06	chr8	130577267	130624661	130077267	131124661
9.01	chr9	4980756	4984530	4480756	5484530

Loci no.	Chromosome	Start locus (bp)	End locus (bp)	Extended start (bp)	Extended end (bp)
9.02	chr9	93904561	93952033	93404561	94452033
9.03	chr9	117538334	117692882	117038334	118192882
9.04	chr9	139257147	139405093	138757147	139905093
10.01	chr10	6038478	6125322	5538478	6625322
10.02	chr10	27160000	27180000	26660000	27680000
10.03	chr10	30689316	30772703	30189316	31272703
10.04	chr10	35256960	35552648	34756960	36052648
10.05	chr10	59901559	60065351	59401559	60565351
10.06	chr10	64348342	64566258	63848342	65066258
10.07	chr10	75469091	75695724	74969091	76195724
10.08	chr10	81032532	81048611	80532532	81548611
10.09	chr10	82214586	82306330	81714586	82806330
10.1	chr10	94248310	94485763	93748310	94985763
10.11	chr10	101274058	101320120	100774058	101820120
10.12	chr10	104217592	104401203	103717592	104901203
10.13	chr10	126320000	126550000	125820000	127050000
11.01	chr11	1873232	1880596	1373232	2380596
11.02	chr11	58174653	58434545	57674653	58934545
11.03	chr11	60776209	60789643	60276209	61289643
11.04	chr11	61543499	61624181	61043499	62124181
11.05	chr11	64133163	64164833	63633163	64664833
11.06	chr11	65575263	65663547	65075263	66163547
11.07	chr11	76281593	76302073	75781593	76802073
11.08	chr11	87011889	87120819	86511889	87620819
11.09	chr11	96018862	96045998	95518862	96545998
11.1	chr11	114323972	114447782	113823972	114947782
11.11	chr11	118758089	118766356	118258089	119266356
11.12	chr11	128380000	128400000	127880000	128900000
12.01	chr12	6490381	6493100	5990381	6993100
12.02	chr12	12613534	12711368	12113534	13211368
12.03	chr12	40214265	40828306	39714265	41328306
12.04	chr12	48195939	48208368	47695939	48708368
12.05	chr12	68476749	68508276	67976749	69008276
12.06	chr12	103410209	113777547	102910209	114277547
12.07	chr12	120146301	120146925	119646301	120646925
13.01	chr13	27531267	27543781	27031267	28043781
13.02	chr13	40678443	41032853	40178443	41532853
13.03	chr13	42840000	42940000	42340000	43440000
13.04	chr13	42951449	43055002	42451449	43555002
13.05	chr13	44406102	44490181	43906102	44990181
13.06	chr13	99778655	100064765	99278655	100564765
14.01	chr14	69254294	69307621	68754294	69807621
14.02	chr14	75702235	75747118	75202235	76247118
14.03	chr14	88404343	88555206	87904343	89055206

Loci no.	Chromosome	Start loci (bp)	End loci (bp)	Extended start (bp)	Extended end (bp)
15.01	chr15	38836777	38925195	38336777	39425195
15.02	chr15	41367036	41687824	40867036	42187824
15.03	chr15	67441750	67468285	66941750	67968285
15.04	chr15	91142885	91221307	90642885	91721307
16.01	chr16	11371759	11718433	10871759	12218433
16.02	chr16	23826417	23867776	23326417	24367776
16.03	chr16	28338043	28895130	27838043	29395130
16.04	chr16	30469919	30514723	29969919	31014723
16.05	chr16	49321282	50827809	48821282	51327809
16.06	chr16	68554754	68680903	68054754	69180903
16.07	chr16	81910000	81920000	81410000	82420000
16.08	chr16	82870000	82920000	82370000	83420000
16.09	chr16	85995436	86011337	85495436	86511337
17.01	chr17	25819513	25869033	25319513	26369033
17.02	chr17	32567679	32640025	32067679	33140025
17.03	chr17	37903731	38089717	37403731	38589717
17.04	chr17	40492540	40546652	39992540	41046652
17.05	chr17	54858402	54949047	54358402	55449047
17.06	chr17	57801597	58046076	57301597	58546076
17.07	chr17	70636731	70642923	70136731	71142923
17.08	chr17	76645300	76858539	76145300	77358539
18.01	chr18	12774326	12818922	12274326	13318922
18.02	chr18	46395022	46395022	45895022	46895022
18.03	chr18	56876228	56893396	56376228	57393396
18.04	chr18	67511645	67546090	67011645	68046090
18.05	chr18	77183529	77237142	76683529	77737142
19.01	chr19	1106477	1127981	606477	1627981
19.02	chr19	10412409	10602180	9912409	11102180
19.03	chr19	33731379	33735149	33231379	34235149
19.04	chr19	46847901	47146676	46347901	47646676
19.05	chr19	49168942	49248730	48668942	49748730
19.06	chr19	55368865	55386920	54868865	55886920
20.01	chr20	6080000	6100000	5580000	6600000
20.02	chr20	30696392	31420757	30196392	31920757
20.03	chr20	33799280	33882720	33299280	34382720
20.04	chr20	43068996	43072706	42568996	43572706
20.05	chr20	44680853	44749251	44180853	45249251
20.06	chr20	48955424	48968438	48455424	49468438
20.07	chr20	57809343	57829301	57309343	58329301
20.08	chr20	62301795	62376939	61801795	62876939
21.01	chr21	16817311	16838662	16317311	17338662
21.02	chr21	34768097	34776695	34268097	35276695
21.03	chr21	40458722	40468838	39958722	40968838
21.04	chr21	45611686	45633388	45111686	46133388

Loci no.	Chromosome	Start locus (bp)	End locus (bp)	Extended start (bp)	Extended end (bp)
22.01	chr22	21911220	21998833	21411220	22498833
22.02	chr22	30269907	30592487	29769907	31092487
22.03	chr22	35720000	35740000	35220000	36240000
22.04	chr22	37260000	37269000	36760000	37769000
22.05	chr22	39659773	39756650	39159773	40256650
22.06	chr22	41648502	42336172	41148502	42836172

Appendix 5 – Differently expressed genes (q ≤ 0.05)

A. CD *versus* control analysis

Gene Name	IBD locus	Extended start (bp)	Extended end (bp)	LogFC	Q-value	Mean FPKM
GLS	2.14	191407655	192472789	-0.42	3.5E-07	99.8
PAPD5	16.05	48821282	51327809	-0.26	7.7E-05	12.8
CORO1C	12.06	102910209	114277547	-0.32	3.7E-04	56.4
PLEKHH3	17.04	39992540	41046652	0.59	3.7E-04	1.6
MTMR2	11.09	95518862	96545998	-0.33	4.7E-04	20.8
HSPE1	2.15	197744598	199454831	-0.29	1.8E-03	99.3
MIER1	1.07	67098347	68243552	-0.26	1.8E-03	80.1
FTL	19.05	48668942	49748730	-0.29	1.8E-03	819.2
RAPGEF3	12.04	47695939	48708368	0.51	1.9E-03	3.6
NOTCH4	6.08	32126272	33126952	0.56	2.1E-03	1.3
SNX13	7.03	16930004	17945706	-0.26	2.4E-03	53.3
ZC3H15	2.13	187000000	188180000	-0.20	2.8E-03	64.4
CAST	5.08	95700770	96873750	-0.24	3.1E-03	313.5
TCTN1	12.06	102910209	114277547	0.31	3.4E-03	7.4
EGFL7	9.04	138757147	139905093	0.55	3.8E-03	3.4
GLTP	12.06	102910209	114277547	-0.25	3.9E-03	39.9
VANGL2	1.17	160337622	161419496	0.84	4.3E-03	1.6
ANO7	2.23	241970000	242990000	0.51	4.4E-03	12.7
ZRANB2	1.08	70491829	71540166	-0.18	4.7E-03	71.2
MOB4	2.15	197744598	199454831	-0.19	4.9E-03	27.8
GRB10	7.08	49596251	50823456	0.36	4.9E-03	5.2
RAPGEF6	5.09	129223552	132333599	-0.22	5.2E-03	23.5
RGS12	4.01	2898068	3950541	0.33	5.2E-03	6.3
GCG	2.11	162610536	163624051	0.89	5.3E-03	51.3
CAPN10	2.22	241063739	242108453	0.25	5.6E-03	11.3
ARFGAP1	20.08	61801795	62876939	0.25	5.8E-03	17.4
UBE2V2	8.02	48547317	49706630	-0.17	6.7E-03	39
P4HTM	3.03	47946237	51595279	0.23	6.9E-03	14.1
GGT7	20.03	33299280	34382720	0.32	7.0E-03	4.5
TAGLN2	1.16	159300000	160390000	-0.26	7.3E-03	194
DLL4	15.02	40867036	42187824	0.36	7.4E-03	5.7
ELF1	13.02	40178443	41532853	-0.19	7.4E-03	48
YIPF2	19.02	9912409	11102180	0.26	7.4E-03	17.5
FCAMR	1.26	206439904	207468955	-0.51	7.6E-03	1.9
ROR2	9.02	93404561	94452033	0.51	7.9E-03	1.7
RNY1	7.15	147900000	149080000	0.69	8.1E-03	160.1
UGT1A1	2.21	233643048	234708258	-0.50	9.1E-03	20.9
UBA7	3.03	47946237	51595279	0.34	9.3E-03	37.1
CHST12	7.01	2252152	3412928	0.50	1.1E-02	2.1
PPTC7	12.06	102910209	114277547	-0.24	1.1E-02	12.3
SNORA74B	5.15	171813034	172829734	0.44	1.1E-02	420.8

Gene Name	IBD locus	Extended start (bp)	Extended end (bp)	LogFC	Q-value	Mean FPKM
RORC	1.14	151292984	152302356	0.35	1.1E-02	11.5
MIB2	1.01	694804	1846703	0.39	1.2E-02	11.5
DNAH17	17.08	76145300	77358539	0.38	1.2E-02	0.8
FER1L4	20.03	33299280	34382720	0.71	1.2E-02	4.4
ASCC2	22.02	29769907	31092487	-0.18	1.2E-02	34.9
GIPC2	1.09	77950517	79123626	-0.21	1.2E-02	25.2
ACAP3	1.01	694804	1846703	0.32	1.2E-02	8.2
SDR42E1	16.07	81410000	82420000	-0.40	1.3E-02	12.8
RP11-81H14.2	12.05	67976749	69008276	0.81	1.3E-02	4.1
FLVCR2	14.02	75202235	76247118	-0.33	1.3E-02	17.1
THBS3	1.15	155112197	156511444	0.33	1.4E-02	5.2
VPS25	17.04	39992540	41046652	-0.19	1.4E-02	58.5
PABPC1L	20.04	42568996	43572706	0.39	1.4E-02	5.3
DENND1B	1.22	196842380	198313558	-0.19	1.4E-02	48.4
ITSN1	21.02	34268097	35276695	-0.19	1.5E-02	18
PITX1	5.1	133922204	134953814	0.56	1.5E-02	4.4
RAC1	7.02	6000000	7050000	-0.20	1.5E-02	179.6
MGAT3	22.05	39159773	40256650	0.53	1.6E-02	2.4
VWA1	1.01	694804	1846703	0.45	1.6E-02	7.7
CAPN10-AS1	2.22	241063739	242108453	0.31	1.6E-02	1.1
RNF215	22.02	29769907	31092487	0.31	1.6E-02	2.4
CCNL2	1.01	694804	1846703	0.19	1.6E-02	54.8
TRAF3IP2	6.12	110993953	112419424	-0.23	1.6E-02	41.7
ZNF282	7.15	147900000	149080000	0.21	1.6E-02	8
TMEM138	11.03	60276209	61289643	-0.17	1.6E-02	28.9
PPP1CB	2.03	28102911	29147084	-0.20	1.6E-02	138.7
MACC1	7.04	20080000	21089000	-0.31	1.7E-02	12.7
SNRNP70	19.05	48668942	49748730	0.25	1.7E-02	48.2
FUBP1	1.09	77950517	79123626	-0.14	1.7E-02	88.4
SEC14L2	22.02	29769907	31092487	0.30	1.7E-02	3.9
GDPD3	16.04	29969919	31014723	-0.47	1.8E-02	40.2
CCDC88B	11.05	63633163	64664833	0.56	1.8E-02	8.8
MAGI3	1.12	113803808	114877568	-0.20	1.8E-02	42.6
MUC1	1.15	155112197	156511444	0.63	1.8E-02	51.9
LYRM7	5.09	129223552	132333599	-0.18	1.8E-02	12.4
PRICKLE4	6.09	41500000	42510000	0.31	1.9E-02	3.5
SLC12A9	7.1	99901433	100933794	0.23	1.9E-02	8.3
RPL26L1	5.16	172769956	173899325	-0.25	1.9E-02	21.9
HAS3	16.06	68054754	69180903	-0.71	2.0E-02	4.5
HOTTIP	7.05	26194926	27411904	0.48	2.0E-02	8.5
MAPRE1	20.02	30196392	31920757	-0.16	2.0E-02	44.8
SAR1B	5.1	133922204	134953814	-0.15	2.1E-02	71.1
SNORA42	7.02	6000000	7050000	0.63	2.1E-02	86.6
CISD1	10.05	59401559	60565351	-0.21	2.1E-02	34.4

Gene Name	IBD locus	Extended start (bp)	Extended end (bp)	LogFC	Q-value	Mean FPKM
TRIM25	17.05	54358402	55449047	-0.23	2.1E-02	30
SLC25A28	10.11	100774058	101820120	0.22	2.1E-02	15.3
SLC4A10	2.11	162610536	163624051	-0.65	2.1E-02	6
UGT1A13P	2.21	233643048	234708258	-0.49	2.1E-02	6.9
GLT8D1	3.04	52478418	53642980	0.16	2.1E-02	25.6
HELZ2	20.08	61801795	62876939	0.48	2.1E-02	5.1
KRTAP5-AS1	11.01	1373232	2380596	0.51	2.2E-02	1.7
ACTR2	2.07	65104914	66192016	-0.19	2.2E-02	122.4
DLD	7.11	106937613	108084780	-0.15	2.2E-02	103.4
DUSP22	6.01	0	903799	0.22	2.2E-02	14.3
ABCA7	19.01	606477	1627981	0.40	2.3E-02	5.3
TP53INP2	20.03	33299280	34382720	-0.43	2.3E-02	33.2
SBNO2	19.01	606477	1627981	0.34	2.4E-02	5.4
CBX7	22.05	39159773	40256650	0.25	2.5E-02	6.4
NGRN	15.04	90642885	91721307	0.21	2.5E-02	46.4
AHCYL2	7.13	128067032	129081835	-0.26	2.6E-02	272.1
LAMB2	3.03	47946237	51595279	0.28	2.6E-02	21.3
FKRP	19.04	46347901	47646676	0.25	2.6E-02	1.8
TNFRSF18	1.01	694804	1846703	0.78	2.7E-02	1.5
CDC42EP5	19.06	54868865	55886920	0.30	2.7E-02	111.4
SERBP1	1.07	67098347	68243552	-0.13	2.8E-02	112.3
USP2	11.11	118258089	119266356	-0.52	2.9E-02	11.5
AFF4	5.09	129223552	132333599	-0.19	2.9E-02	43.3
IDE	10.1	93748310	94985763	-0.18	2.9E-02	34.8
MASTL	10.02	26660000	27680000	-0.23	2.9E-02	7.4
AS3MT	10.12	103717592	104901203	0.35	2.9E-02	1
NISCH	3.04	52478418	53642980	0.21	2.9E-02	17.7
C1orf86	1.02	1970681	3014575	0.27	2.9E-02	5.4
GNG12	1.07	67098347	68243552	-0.24	2.9E-02	107.1
PQLC1	18.05	76683529	77737142	0.25	2.9E-02	15.7
NT5C2	10.12	103717592	104901203	-0.20	2.9E-02	87.7
TMEM135	11.08	86511889	87620819	-0.21	2.9E-02	55.6
PDGFRB	5.12	149090000	150130000	0.36	2.9E-02	7.7
UGT1A10	2.21	233643048	234708258	-0.37	2.9E-02	60.1
KRT8P46	4.08	102891275	104048216	0.36	3.0E-02	1.6
AGRN	1.01	694804	1846703	0.41	3.0E-02	5.4
ALDH1L2	12.06	102910209	114277547	0.42	3.0E-02	2.7
SLC35C2	20.05	44180853	45249251	0.15	3.0E-02	20.9
MAML2	11.09	95518862	96545998	-0.17	3.0E-02	14.3
PSME3	17.04	39992540	41046652	-0.18	3.1E-02	28.2
CGN	1.14	151292984	152302356	-0.26	3.1E-02	76.4
APOM	6.07	30736467	31813602	-0.30	3.1E-02	3.6
FCER1G	1.18	160963601	161979745	-0.34	3.1E-02	38.9
FAM213A	10.09	81714586	82806330	-0.34	3.1E-02	51

Gene Name	IBD locus	Extended start (bp)	Extended end (bp)	LogFC	Q-value	Mean FPKM
SOX9-AS1	17.07	70136731	71142923	0.58	3.1E-02	1.8
INSL5	1.07	67098347	68243552	1.93	3.1E-02	22.1
LENG8	19.06	54868865	55886920	0.23	3.1E-02	40.7
BAD	11.05	63633163	64664833	0.35	3.1E-02	27.8
AC074117.10	2.02	27098097	28252871	0.23	3.2E-02	5.2
RBM6	3.03	47946237	51595279	0.18	3.3E-02	47
AC011298.2	2.22	241063739	242108453	0.86	3.3E-02	0.6
UGT1A6	2.21	233643048	234708258	-0.48	3.3E-02	10.4
RP11-474B16.1	12.06	102910209	114277547	-0.56	3.3E-02	5.4
SLC39A8	4.08	102891275	104048216	0.25	3.3E-02	45.5
CDC42SE2	5.09	129223552	132333599	-0.14	3.4E-02	79.7
NUAK1	12.06	102910209	114277547	0.36	3.4E-02	1.6
RP11-510J16.5	16.07	81410000	82420000	-0.50	3.4E-02	1.9
TNFRSF14	1.02	1970681	3014575	0.20	3.4E-02	48.5
ATP5E	20.07	57309343	58329301	-0.17	3.5E-02	53
NRCAM	7.11	106937613	108084780	0.63	3.6E-02	1
ARID3A	19.01	606477	1627981	0.37	3.6E-02	1.3
SEH1L	18.01	12274326	13318922	-0.17	3.6E-02	17.5
UGT1A7	2.21	233643048	234708258	-0.47	3.6E-02	8.8
MEI1	22.06	41148502	42836172	0.51	3.6E-02	9.8
CD244	1.17	160337622	161419496	-0.44	3.7E-02	1.8
DVL1	1.01	694804	1846703	0.25	3.7E-02	10.6
GDPGP1	15.04	90642885	91721307	-0.24	3.8E-02	8.8
KIFAP3	1.19	168590748	170019049	-0.23	3.9E-02	33.3
RABEP2	16.03	27838043	29395130	0.24	4.0E-02	13.8
FAM193B	5.17	176282218	177306636	0.21	4.0E-02	17.7
SUV420H2	19.06	54868865	55886920	0.34	4.0E-02	3.4
LAD1	1.25	200374229	201524059	-0.24	4.0E-02	75.6
IPMK	10.05	59401559	60565351	-0.20	4.0E-02	11.7
C12orf23	12.06	102910209	114277547	0.25	4.1E-02	16.7
HCG27	6.07	30736467	31813602	0.41	4.1E-02	2.8
FZD8	10.04	34756960	36052648	0.38	4.2E-02	1.6
PPM1F	22.01	21411220	22498833	0.23	4.2E-02	3.7
BRAT1	7.01	2252152	3412928	0.27	4.2E-02	7.2
CFD	19.01	606477	1627981	0.41	4.2E-02	16
EMILIN1	2.02	27098097	28252871	0.39	4.3E-02	6.8
CHTF8	16.06	68054754	69180903	-0.16	4.3E-02	32.6
ANKRD16	10.01	5538478	6625322	0.22	4.3E-02	3.2
DYRK2	12.05	67976749	69008276	-0.22	4.4E-02	51.3
IFT81	12.06	102910209	114277547	0.21	4.4E-02	4.8
HSPA7	1.18	160963601	161979745	0.77	4.5E-02	1.8
UGT1A4	2.21	233643048	234708258	-0.48	4.7E-02	2.5
GPC1	2.22	241063739	242108453	0.32	4.7E-02	2.2

Gene Name	IBD locus	Extended start (bp)	Extended end (bp)	LogFC	Q-value	Mean FPKM
DNAJB3	2.21	233643048	234708258	-0.53	4.7E-02	3.1
SEMA3F	3.03	47946237	51595279	0.43	4.7E-02	2.6
MLX	17.04	39992540	41046652	-0.17	4.8E-02	34.3
C1QB	1.05	22181214	23211473	-0.32	4.8E-02	98.4
CHP1	15.02	40867036	42187824	-0.22	4.9E-02	178.4
PFKFB4	3.03	47946237	51595279	0.31	4.9E-02	4.9

B. IBD *versus* control analysis

Gene Name	IBD locus	Extended start (bp)	Extended end (bp)	LogFC	q-value	Mean FPKM
GLS	2.14	191407655	192472789	-0.32	8.4E-05	99.8
MTMR2	11.09	95518862	96545998	-0.31	8.9E-04	20.8
HSPE1	2.15	197744598	199454831	-0.28	1.6E-03	99.3
RAPGEF6	5.09	129223552	132333599	-0.23	1.7E-03	23.5
PAPD5	16.05	48821282	51327809	-0.21	1.7E-03	12.8
NOTCH4	6.08	32126272	33126952	0.48	2.0E-03	1.3
MOB4	2.15	197744598	199454831	-0.19	2.2E-03	27.8
ANO7	2.23	241970000	242990000	0.50	2.7E-03	12.7
MIER1	1.07	67098347	68243552	-0.23	4.1E-03	80.1
TRIM25	17.05	54358402	55449047	-0.25	6.1E-03	30
RNY1	7.15	147900000	149080000	0.71	6.2E-03	160.1
FTL	19.05	48668942	49748730	-0.23	6.6E-03	819.2
ZRANB2	1.08	70491829	71540166	-0.17	6.8E-03	71.2
CORO1C	12.06	102910209	114277547	-0.23	8.7E-03	56.4
PPTC7	12.06	102910209	114277547	-0.22	1.0E-02	12.3
SERBP1	1.07	67098347	68243552	-0.14	1.1E-02	112.3
RGS12	4.01	2898068	3950541	0.27	1.1E-02	6.3
DNAH17	17.08	76145300	77358539	0.36	1.1E-02	0.8
FUBP1	1.09	77950517	79123626	-0.14	1.1E-02	88.4
TAGLN2	1.16	159300000	160390000	-0.23	1.2E-02	194
GRB10	7.08	49596251	50823456	0.30	1.2E-02	5.2
IFT81	12.06	102910209	114277547	0.22	1.4E-02	4.8
VANGL2	1.17	160337622	161419496	0.69	1.4E-02	1.6
EGFL7	9.04	138757147	139905093	0.45	1.6E-02	3.4
ZC3H15	2.13	187000000	188180000	-0.16	1.6E-02	64.4
SOX9-AS1	17.07	70136731	71142923	0.60	1.6E-02	1.8
AGPAT2	9.04	138757147	139905093	-0.33	1.7E-02	78.6
FCER1G	1.18	160963601	161979745	-0.33	1.8E-02	38.9
OSGIN2	8.03	90354846	91377546	-0.17	1.8E-02	6.5
KRTAP5-AS1	11.01	1373232	2380596	0.47	1.8E-02	1.7
UBE2V2	8.02	48547317	49706630	-0.15	1.9E-02	39
C1QB	1.05	22181214	23211473	-0.35	1.9E-02	98.4
HLA-DRB5	6.08	32126272	33126952	2.63	1.9E-02	22.4
CDH13	16.08	82370000	83420000	0.51	1.9E-02	7.1
NXPE1	11.10	113823972	114947782	0.29	2.1E-02	199.8
KCNH8	3.01	18199977	19325669	0.40	2.2E-02	1
TCTN1	12.06	102910209	114277547	0.25	2.3E-02	7.4
FAM49B	8.06	130077267	131124661	-0.15	2.4E-02	41.8
IQCH	15.03	66941750	67968285	0.29	2.6E-02	2.4
GLTP	12.06	102910209	114277547	-0.20	2.6E-02	39.9
EMC8	16.09	85495436	86511337	-0.20	2.7E-02	9.1
PRICKLE4	6.09	41500000	42510000	0.28	2.7E-02	3.5

Gene Name	IBD locus	Extended start (bp)	Extended end (bp)	LogFC	Q-value	Mean FPKM
RBM5	3.03	47946237	51595279	0.12	2.9E-02	127.5
SNORA74B	5.15	171813034	172829734	0.38	2.9E-02	420.8
RNF26	11.11	118258089	119266356	-0.25	2.9E-02	5.3
RORC	1.14	151292984	152302356	0.33	2.9E-02	11.5
CD244	1.17	160337622	161419496	-0.44	2.9E-02	1.8
UBA7	3.03	47946237	51595279	0.28	2.9E-02	37.1
PPP1CB	2.03	28102911	29147084	-0.18	2.9E-02	138.7
ACTR2	2.07	65104914	66192016	-0.18	3.0E-02	122.4
RBM6	3.03	47946237	51595279	0.17	3.0E-02	47
MASTL	10.02	26660000	27680000	-0.22	3.0E-02	7.4
ARHGEF28	5.07	72002029	73059339	0.25	3.4E-02	10
VPS25	17.04	39992540	41046652	-0.17	3.4E-02	58.5
HMBS	11.11	118258089	119266356	-0.25	3.4E-02	9.6
CAST	5.08	95700770	96873750	-0.17	3.5E-02	313.5
ELF1	13.02	40178443	41532853	-0.14	3.5E-02	48
MAPRE1	20.02	30196392	31920757	-0.15	3.5E-02	44.8
DLL4	15.02	40867036	42187824	0.27	3.6E-02	5.7
ASCC2	22.02	29769907	31092487	-0.14	3.7E-02	34.9
AS3MT	10.12	103717592	104901203	0.30	3.8E-02	1
LAD1	1.25	200374229	201524059	-0.23	3.8E-02	75.6
ITLN1	1.17	160337622	161419496	0.42	3.9E-02	389.8
CTSZ	20.07	57309343	58329301	-0.18	4.0E-02	104.3
SNORA42	7.02	6000000	7050000	0.54	4.0E-02	86.6
SEC14L2	22.02	29769907	31092487	0.24	4.1E-02	3.9
PFKFB4	3.03	47946237	51595279	0.28	4.2E-02	4.9
RAPGEF3	12.04	47695939	48708368	0.35	4.2E-02	3.6
CHTF8	16.06	68054754	69180903	-0.14	4.2E-02	32.6
FAM65C	20.06	48455424	49468438	0.28	4.3E-02	3
VIL1	2.18	218566980	219691569	-0.19	4.5E-02	185.4
GDPD3	16.04	29969919	31014723	-0.39	4.6E-02	40.2
FCAMR	1.26	206439904	207468955	-0.38	4.6E-02	1.9
PCNP	3.07	100410000	101770000	-0.12	4.6E-02	39.5
PSME3	17.04	39992540	41046652	-0.16	4.8E-02	28.2
CACNA2D2	3.03	47946237	51595279	0.35	4.9E-02	1
CFAP70	10.07	74969091	76195724	0.31	4.9E-02	1.7
P4HTM	3.03	47946237	51595279	0.17	4.9E-02	14.1
CSTB	21.04	45111686	46133388	-0.32	4.9E-02	37.4
HCG27	6.07	30736467	31813602	0.37	5.0E-02	2.8

C. UC *versus* CD analysis

Gene Name	IBD locus	Extended start (bp)	Extended end (bp)	LogFC	Q-value	Mean FPKM
TMEM259	19.01	606477	1627981	-0.36	4.01E-05	21.8
GAL3ST2	2.23	241970000	242990000	-0.96	2.39E-04	4.3
GPC1	2.22	241063739	242108453	-0.56	4.72E-04	2.2
FER1L4	20.03	33299280	34382720	-0.99	5.22E-04	4.4
BAD	11.05	63633163	64664833	-0.52	5.22E-04	27.8
MBD3	19.01	606477	1627981	-0.33	1.03E-03	11
HMHA1	19.01	606477	1627981	-0.51	1.05E-03	7.7
YIPF2	19.02	9912409	11102180	-0.31	1.53E-03	17.5
CHST12	7.01	2252152	3412928	-0.58	1.72E-03	2.1
PAPD5	16.05	48821282	51327809	0.23	1.75E-03	12.8
LFNG	7.01	2252152	3412928	-0.40	2.04E-03	18.4
AGRN	1.01	694804	1846703	-0.52	4.30E-03	5.4
H2AFX	11.11	118258089	119266356	-0.52	4.56E-03	11.9
LRP3	19.03	33231379	34235149	-0.49	5.03E-03	1.8
ZBTB46	20.08	61801795	62876939	-0.42	5.33E-03	1.8
SOCS6	18.04	67011645	68046090	0.27	5.35E-03	25.9
LRCH4	7.1	99901433	100933794	-0.30	5.35E-03	15.4
SLC2A4RG	20.08	61801795	62876939	-0.38	5.45E-03	13.1
ZNF512B	20.08	61801795	62876939	-0.56	5.74E-03	0.8
RAPGEF3	12.04	47695939	48708368	-0.51	6.02E-03	3.6
NDUFS7	19.01	606477	1627981	-0.42	6.02E-03	61.5
SMPD3	16.06	68054754	69180903	-0.34	6.31E-03	19.5
FBXL15	10.12	103717592	104901203	-0.53	6.68E-03	3.5
NUDT22	11.05	63633163	64664833	-0.35	6.68E-03	44.1
LRPAP1	4.01	2898068	3950541	-0.20	6.68E-03	49.8
MIB2	1.01	694804	1846703	-0.41	7.03E-03	11.5
PLEKHA4	19.05	48668942	49748730	-0.49	8.25E-03	0.9
SLC1A5	19.04	46347901	47646676	-0.32	8.73E-03	32.2
RAC1	7.02	6000000	7050000	0.24	9.21E-03	179.6
AC114730.3	2.24	242224543	243199373	-0.70	9.21E-03	3.1
DVL1	1.01	694804	1846703	-0.32	9.21E-03	10.6
CFD	19.01	606477	1627981	-0.59	9.21E-03	16
STMN3	20.08	61801795	62876939	-0.52	9.21E-03	2.2
UGT1A10	2.21	233643048	234708258	0.49	1.01E-02	60.1
DENND1B	1.22	196842380	198313558	0.22	1.02E-02	48.4
GLS	2.14	191407655	192472789	0.25	1.08E-02	99.8
CBX7	22.05	39159773	40256650	-0.31	1.08E-02	6.4
UGT1A13P	2.21	233643048	234708258	0.66	1.09E-02	6.9
ACAP3	1.01	694804	1846703	-0.35	1.09E-02	8.2
ABHD16B	20.08	61801795	62876939	-0.55	1.09E-02	1.2
RGS14	5.17	176282218	177306636	-0.53	1.20E-02	6
RPS6KB1	17.06	57301597	58546076	0.15	1.20E-02	12.6
FBXW5	9.04	138757147	139905093	-0.27	1.25E-02	45.6

Gene Name	IBD locus	Extended start (bp)	Extended end (bp)	LogFC	Q-value	Mean FPKM
PQLC1	18.05	76683529	77737142	-0.28	1.26E-02	15.7
CCDC85B	11.06	65075263	66163547	-0.79	1.33E-02	1.6
P4HTM	3.03	47946237	51595279	-0.24	1.33E-02	14.1
ATP5D	19.01	606477	1627981	-0.36	1.35E-02	60.7
ACER3	11.07	75781593	76802073	0.25	1.35E-02	31.2
IQCD	12.06	102910209	114277547	0.58	1.35E-02	1.6
RAB24	5.17	176282218	177306636	-0.45	1.35E-02	11.1
SNRNP70	19.05	48668942	49748730	-0.26	1.36E-02	48.2
GFI1	1.1	92054283	93054283	-0.30	1.36E-02	5.1
CDC42EP5	19.06	54868865	55886920	-0.35	1.39E-02	111.4
USP32	17.06	57301597	58546076	0.17	1.54E-02	11.3
ARFGAP1	20.08	61801795	62876939	-0.24	1.58E-02	17.4
RNASET2	6.19	166860389	167985800	-0.24	1.58E-02	72.2
FLVCR2	14.02	75202235	76247118	0.37	1.69E-02	17.1
RPL17P11	2.21	233643048	234708258	0.63	1.76E-02	3
PLEKHH3	17.04	39992540	41046652	-0.43	1.76E-02	1.6
MACC1	7.04	20080000	21089000	0.32	1.76E-02	12.7
IRF5	7.13	128067032	129081835	-0.40	1.76E-02	4.1
LENG9	19.06	54868865	55886920	-0.40	1.84E-02	4.7
TMEM160	19.04	46347901	47646676	-0.64	1.97E-02	3.7
PPM1F	22.01	21411220	22498833	-0.30	1.98E-02	3.7
LEPREL2	12.01	5990381	6993100	-0.48	2.04E-02	1.8
UGT1A1	2.21	233643048	234708258	0.52	2.06E-02	20.9
CAST	5.08	95700770	96873750	0.21	2.16E-02	313.5
LPP	3.11	187900000	188990000	0.19	2.20E-02	151.6
C1QTNF6	22.04	36760000	37769000	-0.36	2.24E-02	3.6
YOD1	1.26	206439904	207468955	0.24	2.29E-02	7.9
INSL5	1.07	67098347	68243552	-2.11	2.32E-02	22.1
AC009506.1	2.1	160191494	161378364	-0.35	2.39E-02	5.4
GAL3ST1	22.02	29769907	31092487	-0.80	2.45E-02	2.3
NT5C2	10.12	103717592	104901203	0.22	2.45E-02	87.7
MEI1	22.06	41148502	42836172	-0.58	2.54E-02	9.8
SPAG4	20.03	33299280	34382720	-0.61	2.54E-02	3.1
C2orf82	2.21	233643048	234708258	-0.28	2.54E-02	7.7
NEU4	2.23	241970000	242990000	-0.83	2.58E-02	10.7
CD27	12.01	5990381	6993100	-0.63	2.68E-02	9.7
PUSL1	1.01	694804	1846703	-0.36	2.68E-02	5.4
SLC12A9	7.1	99901433	100933794	-0.23	2.70E-02	8.3
ZRANB2	1.08	70491829	71540166	0.16	2.72E-02	71.2
CKAP4	12.06	102910209	114277547	-0.34	2.77E-02	26.1
GNG12	1.07	67098347	68243552	0.27	2.78E-02	107.1
PIN1	19.02	9912409	11102180	-0.21	2.78E-02	23.5
FAM193B	5.17	176282218	177306636	-0.24	2.86E-02	17.7
PRKAA1	5.04	39719972	41123346	0.14	2.86E-02	75.3

Gene Name	IBD locus	Extended start (bp)	Extended end (bp)	LogFC	Q-value	Mean FPKM
C9orf91	9.03	117038334	118192882	-0.28	2.88E-02	4.4
SLC41A2	12.06	102910209	114277547	0.24	2.93E-02	85
PPP1R35	7.1	99901433	100933794	-0.33	3.01E-02	8.4
C1orf86	1.02	1970681	3014575	-0.30	3.03E-02	5.4
GCG	2.11	162610536	163624051	-0.79	3.15E-02	51.3
PEX10	1.02	1970681	3014575	-0.18	3.15E-02	10.8
GLTPD1	1.01	694804	1846703	-0.31	3.15E-02	13.9
APOBEC3C	22.05	39159773	40256650	-0.49	3.23E-02	8.1
AFF4	5.09	129223552	132333599	0.19	3.30E-02	43.3
GHDC	17.04	39992540	41046652	-0.23	3.30E-02	6.6
MAML2	11.09	95518862	96545998	0.18	3.44E-02	14.3
TAS1R3	1.01	694804	1846703	-0.53	3.48E-02	0.8
YDJC	22.01	21411220	22498833	-0.30	3.48E-02	7.1
BRAT1	7.01	2252152	3412928	-0.29	3.51E-02	7.2
CGREF1	2.02	27098097	28252871	-0.91	3.56E-02	1.9
SNX8	7.01	2252152	3412928	-0.33	3.58E-02	11
MXD3	5.17	176282218	177306636	-0.36	3.58E-02	18.5
C9orf142	9.04	138757147	139905093	-0.30	3.62E-02	16.5
SAR1B	5.1	133922204	134953814	0.15	3.64E-02	71.1
CDX1	5.12	149090000	150130000	-0.21	3.73E-02	122.2
CAMTA1	1.03	7469507	8686232	0.18	3.73E-02	45
CARD9	9.04	138757147	139905093	-0.52	3.80E-02	1.3
CAPN10	2.22	241063739	242108453	-0.21	3.83E-02	11.3
WNK4	17.04	39992540	41046652	-0.43	3.88E-02	6.5
BCL7C	16.04	29969919	31014723	-0.22	3.91E-02	16.1
GGT7	20.03	33299280	34382720	-0.29	4.08E-02	4.5
HLA-DRB6	6.08	32126272	33126952	-2.13	4.08E-02	0.9
B3GALT5	21.03	39958722	40968838	-0.73	4.08E-02	35.1
UGT1A7	2.21	233643048	234708258	0.53	4.26E-02	8.8
UGT1A4	2.21	233643048	234708258	0.55	4.26E-02	2.5
PTMS	12.01	5990381	6993100	-0.25	4.30E-02	25.9
MRPL23	11.01	1373232	2380596	-0.21	4.30E-02	51.8
DEXI	16.01	10871759	12218433	-0.33	4.42E-02	7.4
FAM83E	19.05	48668942	49748730	-0.27	4.42E-02	16.5
TMEM135	11.08	86511889	87620819	0.22	4.44E-02	55.6
TNFRSF14	1.02	1970681	3014575	-0.19	4.45E-02	48.5
RP1-170O19.14	7.06	26731762	27748891	-0.76	4.45E-02	2.6
RP11-119F19.2	10.08	80532532	81548611	0.32	4.52E-02	3.3
ADCY7	16.05	48821282	51327809	-0.30	4.55E-02	9.2
RGS19	20.08	61801795	62876939	-0.41	4.55E-02	4.3
HLA-DPB1	6.08	32126272	33126952	-0.44	4.59E-02	60.4
UGT1A9	2.21	233643048	234708258	0.51	4.67E-02	6.4
UGT1A8	2.21	233643048	234708258	0.59	4.69E-02	5.9

Gene Name	IBD locus	Extended start (bp)	Extended end (bp)	LogFC	Q-value	Mean FPKM
RP11-864J10.4	12.06	102910209	114277547	0.28	4.69E-02	2.1
TTC33	5.04	39719972	41123346	0.19	4.79E-02	15.2
CORO1C	12.06	102910209	114277547	0.21	4.83E-02	56.4
NIPAL1	4.05	47844930	48930354	0.20	4.92E-02	32.4

Appendix 6 – Cell count predictors

CD45^{pos} Leukocyte predictive genes	
Ensemble Gene ID	**Estimate value**
ENSG00000002586	-0.000324
ENSG00000003249	-0.00699
ENSG00000003989	-0.000047664
ENSG00000004059	-0.001178
ENSG00000005073	0.000758
ENSG00000005156	-0.001298
ENSG00000005194	0.002613
ENSG00000005238	-0.0017
ENSG00000005486	0.003036
ENSG00000006025	0.000181
ENSG00000006125	0.000189
ENSG00000006534	-0.001169
ENSG00000006555	-0.000385
ENSG00000006712	0.002947
ENSG00000006744	-0.000438
ENSG00000007923	-0.000294
ENSG00000008283	-0.001103
ENSG00000008300	0.008134
ENSG00000009765	-0.000065815
ENSG00000010818	-0.000031697
Intercept	0.62088

CD326^{pos} Epithelial cell predictive gene	
Ensemble Gene ID	**Estimate value**
ENSG00000011009	-0.00011
ENSG00000011083	-3.6E-05
ENSG00000011275	-0.00034
ENSG00000012779	-0.00059
ENSG00000012822	-0.00031
ENSG00000013275	-0.00023
ENSG00000013364	-3.9E-05
ENSG00000013441	0.000186
ENSG00000013503	0.001032
ENSG00000013563	0.00081
ENSG00000014216	0.000184
ENSG00000014257	0.000344
ENSG00000015133	-0.00024
ENSG00000019505	-0.00029
ENSG00000021355	-4.2E-05
ENSG00000023191	0.000228
ENSG00000023445	3.72E-06
ENSG00000023516	0.0002
ENSG00000023608	0.001983
ENSG00000025708	-0.00022
Intercept	0.314533

CD4^{pos} T helper cell predictive genes	
Ensemble Gene ID	**Estimate value**
ENSG00000011009	2.83E-05
ENSG00000011083	-2.2E-05
ENSG00000011275	2.33E-05
ENSG00000012779	7.02E-05
ENSG00000012822	3.61E-06
ENSG00000013275	1.59E-05
ENSG00000013364	8.15E-06
ENSG00000013441	1.94E-06
ENSG00000013503	-0.0002
ENSG00000013563	-0.00019
ENSG00000014216	-2.5E-05
ENSG00000014257	-4E-05
ENSG00000015133	2.8E-05
ENSG00000019505	8.36E-06
ENSG00000021355	1.65E-06
ENSG00000023191	6.61E-06
Intercept	-0.03271

CD8^{pos} Cytotoxic T cell predictive genes	
Ensemble Gene ID	**Estimate value**
ENSG00000011009	-6.7E-06
ENSG00000011083	6.49E-05
ENSG00000011275	5.18E-05
ENSG00000012779	-2.1E-06
ENSG00000012822	-3.5E-05
ENSG00000013275	-2.4E-06
ENSG00000013364	5.06E-06
ENSG00000013441	2.09E-05
ENSG00000013503	5.55E-05
ENSG00000013563	-4.1E-05
ENSG00000014216	0.000004
ENSG00000014257	-9.7E-06
ENSG00000015133	-1.5E-05
ENSG00000019505	-4.7E-05
ENSG00000021355	-4.1E-06
ENSG00000023191	-7.9E-07
Intercept	-0.09313

CD14^{pos} Monocytes predictive genes	
Ensemble Gene ID	**Estimate value**
ENSG00000011009	4.57E-06
ENSG00000011083	-5.9E-05
ENSG00000011275	-3.6E-05
ENSG00000012779	6.45E-05
ENSG00000012822	-1.4E-05
ENSG00000013275	-1.4E-05
ENSG00000013364	2.05E-06
ENSG00000013441	6.44E-06
ENSG00000013503	-8.4E-05
ENSG00000013563	-8.7E-05
ENSG00000014216	-1.7E-05
ENSG00000014257	5.12E-05
ENSG00000015133	2.58E-05
ENSG00000019505	-1E-05
ENSG00000021355	2.84E-06
ENSG00000023191	4.54E-06
Intercept	0.215859